COMPUTATIONAL DETECTION AND ANALYSIS OF TRANSCRIPTIONAL CONTROL ELEMENTS IN LYMPHOCYTE DEVELOPMENT

VANDANA SINGH

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in the School of Informatics
Indiana University

August 2006

Accepted by the graduate faculty of Indiana University in partial fulfillment of the requirements for the degree of Master of Science.

_____

Dr. Narayanan Perumal

_____

Dr. Mark Kaplan

Master's Thesis Committee

_____

Dr. Huanmei Wu

ACKNOWLEDGEMENTS

While completing this thesis, I have been blessed with the opportunity to learn new and exciting areas of research. I am grateful to all individuals who have helped me to complete my thesis work.

I would like to express my heartfelt gratitude to my advisor, Dr. Narayanan Perumal, for his guidance, support, encouragement and excellent advice throughout this thesis.

I would also like to thank both Dr. Mark Kaplan and Dr. Huanmei Wu for their wonderful advice and for being a part of the thesis committee.

I deeply appreciate all the love and support that I have received from my family.

ABSTRACT

BACKGROUND: Lymphocyte development and differentiation in mammals follow complex gene regulatory mechanisms with control at the transcriptional stage playing a major role. B and T cells, the two large subsets of lymphocytes, develop differentially due to the varying expression patterns of a variety of genes. Computational tools and methods are becoming increasingly useful in the elucidation of various mechanisms in this process, which has traditionally been studied by experimentation. Wet laboratory experimentation invariably consists of studying one gene at a time although recent advances in microarray and chromatin immunoprecipitation (ChIP) technologies have made available large data sets for informatics analysis. Another impetus for computational approaches has been the explosion of annotated mammalian genomic data in various databases. Traditionally, DNA sequences upstream of the expressed genes (cis-acting) and transcription factor molecules binding to these DNA sequences (trans-acting) have been explored. We have been employing a computational regimen to identify transcriptional control elements in the DNA (promoters) of genes that may differentiate the development of B and T cells. Towards this goal, our scheme involves the collection and analysis of four different data sets specific for genes involved in B and T cell development with the focus being on the sequences upstream to the transcription start site (TSS) of the relevant genes.

RESULTS: Using datasets of B and T cell specific genes (Immunoglobulin and T cell receptor genes respectively) from RefSeq, we have identified two predominant consensus patterns in their upstream regions using the Gibbs Recursive Sampler software. With the

help of transcription factor binding site (TFBS) prediction software, different TFBS were obtained for B and T cell genes on the same datasets. A few of them are biologically important, for example, in the case of B cell specific genes we obtained Oct-1, a known immune-specific TFBS. We employed MEME and Gibbs Recursive Sampler software on two different data sets of B and T cell specific regulatory sequences and found different motifs, which are carried by genes common in both software predictions and further used the EZ-Retrieve tool on different motifs to find TFBS. We predicted several immunologically relevant TFBS, such as E47, Oct-1 and GATA-1, at different locations and on both strands in these motifs. In addition, k-means clustering was performed on the datasets in order to classify the B and T cell genes based on the frequencies of TFBS in their upstream sequences. Applying several computational methods, we are able to find additional information on B and T cell genes in terms of TFBS, which may help in the understanding of B and T cell development.

CONCLUSIONS: Performing computational approaches like MEME, Gibbs Recursive Sampler, statistical analysis and k-means clustering on different DNA (promoter) sequences does not always identify biologically meaningful transcriptional control elements involved in lymphocyte development. On the other hand, our predictions of conserved motifs in upstream regulatory regions of target genes, and in particular, the identification of immune-specific TFBS in these motifs are biologically relevant. We hope that they will provide a guide for the experimental biologist to focus on certain elements for biological validation. In summary, this informatics approach to detect

transcriptional control elements may efficiently and effectively aid the biologist to study transcriptional regulation that distinguishes B and T cell development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Introduction to the Subject

The study of transcriptional regulation in the eukaryotic genomes is a major challenge in current molecular biology. Although researchers are taking several important steps to understand and identify the regulatory elements, research in this field needs significant advances (1). Regulation of gene expression occurs in the transcriptional process at different levels. This includes initiation of RNA transcription by transcription factors (TFs) that attract the RNA polymerase II complex. Initiation of transcription is probably the most important regulatory step in control of transcriptional regulation. TFs, proteins that bind to specific DNA sequences present in the core promoters and enhancers of genes, are important in the initiation of gene-specific transcription. Before they can exert their action, initiation of transcription is controlled by more global mechanisms that regulate accessibility of a locus.

Biologists have determined that the control of gene expression at the transcriptional level is primarily determined by relatively short sequences called TFBS in the upstream regions of the genes (2). These sequences vary in length, position, redundancy, orientation, and bases. Finding these short sequences is a fundamental problem in molecular biology, which has important applications.

Development of B and T lymphocytes under complex transcriptional regulatory mechanisms has been studied extensively for the past two decades. Development of B and T cells occurs at various stages. Transcriptional regulation occurs at these stages and different transcription factors are expressed. A variety of modern methods are being used to detect TFBS in B and T cells.

There are different computational methods, which are applied on genomic sequences to identify these TFBS. The various computational approaches to detect such TFBS may successfully help the biologist to study transcriptional regulation that distinguishes B and T cell development. In this project we have used several computational methods to identify and analyze the TFBS using upstream sequences of lymphocyte genes from human and mouse species.

1.1.1 General Transcriptional Regulation

Transcription is the process to synthesize RNA from double stranded DNA. In eukaryotic transcription, RNA polymerases I, II and III transcribe rRNA, mRNA and tRNA genes respectively. RNA polymerases can transcribe with the help of transcription factors binding at promoters (Figure 1).

In general, promoters and enhancers are regions upstream of the transcriptional start site (TSS), the nucleotide at which transcription begins. The TFs bind to the TFBS in the promoter or enhancer regions (Figure 1).

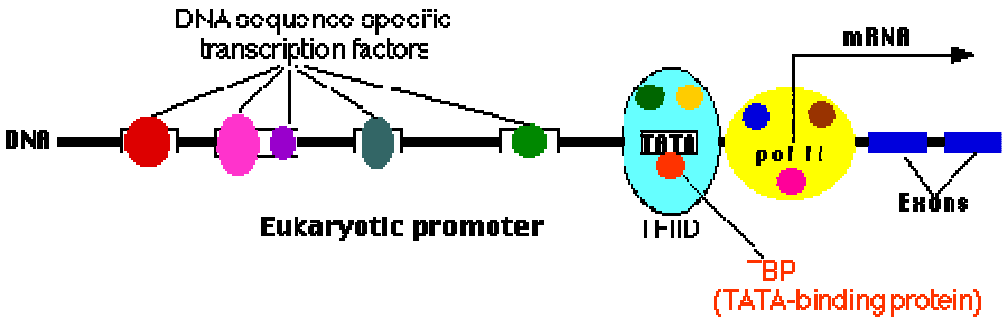


Figure 1. Structure of the transcriptional unit of an eukaryotic mRNA gene (3).

Steps involved in transcription include initiation, elongation and termination. RNA polymerase II recognizes the promoters and unwinds the DNA double helix, and

then initiates transcription of the downstream DNA. RNA polymerase II is associated with several transcription factors, such as TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH. TFIID consists of TATA binding protein (TBP) and TBP associated factors (TAFs). The role of TBP is to bind the core promoter (4). The transcription factor, which catalyzes DNA melting, is TFIIH. However, before TFIIH can unwind DNA, the RNA polymerase II and at least five general transcription factors (TFIIA is not absolutely necessary) have to form a pre-initiation complex (PIC).

After PIC is assembled at the promoter, TFIIH can use its helicase activity to unwind DNA. This requires energy released from ATP hydrolysis. The DNA melting starts from about -10 bp. The RNA polymerase II then uses nucleoside triphosphates (NTPs) to synthesize a RNA transcript. During RNA elongation, TFIIF remains attached to the RNA polymerase, but all of the other transcription factors have dissociated from the PIC.

The carboxyl-terminal domain (CTD) of the largest subunit of RNA polymerase II is critical for elongation. In the initiation phase, CTD is unphosphorylated, but during elongation it has to be phosphorylated. This domain contains many proline, serine and threonine residues.

Eukaryotic protein genes contain a poly-A signal located downstream of the last exon. This signal is used to add a series of adenine residues during RNA processing. Transcription often terminates at 0.5 - 2 kb downstream of the poly-A signal (4).

1.1.2. Lymphocyte Development

It is known that immunology is the study of the body's defenses against infectious microorganisms. Lymphocytes are a major type of cells responsible for generating

immune responses. Lymphocytes are activated by antigens to give clones of antigen specific cells that mediate adaptive immunity. B and T cell, the two large subsets of lymphocytes, develop differentially due to the varying expression patterns of a variety of genes. Lymphocyte development and differentiation in mammals follows complex gene regulatory mechanisms with control at the transcriptional stage playing a major role**.**

All lymphocytes are derived from a stem cell in the bone marrow. T lymphocytes go to the thymus (a large lymphoid organ in the upper chest) for maturation and antigen specificity (5) while B lymphocytes undergo maturation in the bone marrow (6), as shown in Figure 2. Mature lymphocytes migrate from these tissues via blood to peripheral lymphoid organs, organized tissues such as lymph node, spleen and gut-associated lymphoid tissues.

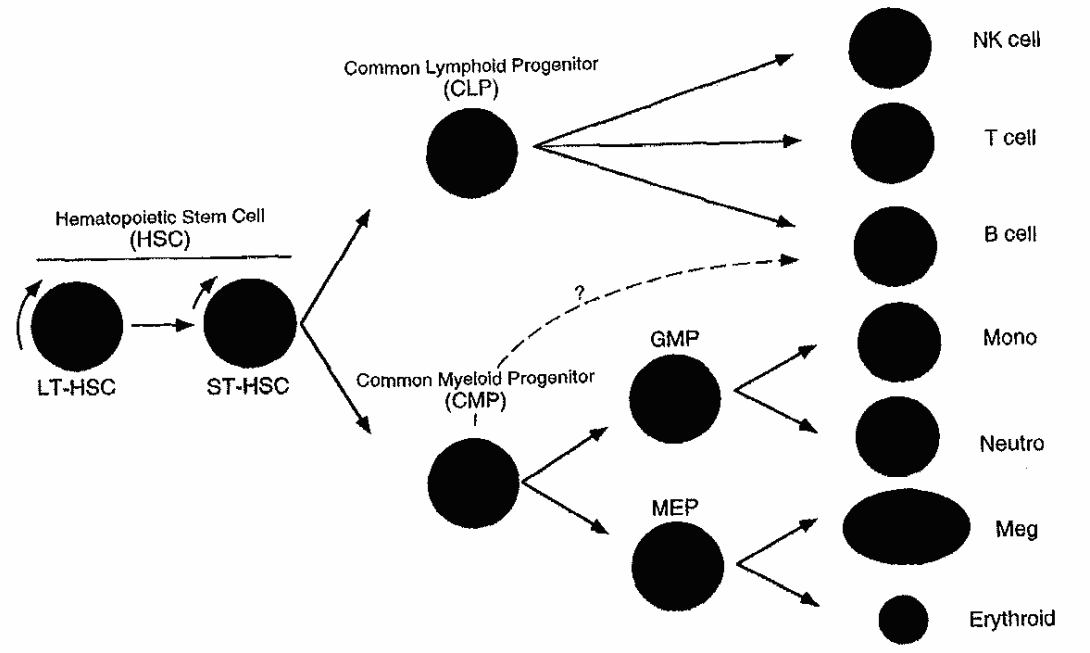


Figure 2. Developmental pathways of lymphoid and myeloid lineages. HSE includes long-term (LT) and short-term (ST) HSCs. Common lymphoid progenitors are able to generate all lymphocytes as NK cells, T cells and B cells; on the other hand common myeloid progenitors give rise to all myeloerythroid cells, megakaryocyte/erythroid restricted progenitors (MEP) and granulocyte/macrophage-restricted progenitors (GMP) (7).

Development of B cells

Development of B cells occurs in bone marrow and is dependent upon bone marrow stromal cells (6). B cell development proceeds through a number of stages with accompanying rearrangements of immunoglobulin genes.

Pre-pro-B cells and B cell Commitment

The initial B cell precursors are pre-pro-B cells. These cells have their immunoglobulin gene loci in germline configuration but they are different from earlier precursors by a series of cell surface markers. Pre-pro-B cells have low expression of the recombination activating genes (RAG-1 and RAG-2) and they do not express components of the B-cell antigen receptor (BCR). Thus, commitment to the B cell pathway precedes antibody (Ab) gene recombination and is BCR-independent. Several transcription factors have been identified as key regulators of B cell development. Among these Pax-5 is unique because it appears to be essential for maintaining B lineage commitment. Pax-$5^{-/-}$ pro-B cells express B lineage markers and initiate immunoglobulin rearrangements but then fail to mature into B cells, instead giving rise to myeloid and T lineage cells. Thus, Pax-5 is essential for B cell commitment and repressing alternative lineage differentiation.

Pro-B-cells and Early Immunoglobulin Gene Recombination

In the pro-B cell V (D)J rearrangement takes place , and it is the first B lineage cell to express a precursor form of the BCR composed of immunoglobulin α (Igα), Igβ and calnexin (pro-BCR) (Figure3). Igα and Igβ are BCR signaling components associated with membrane bound Igμ (mIgμ) in more mature B cells. These immunoglobulin family members activate cellular signaling pathways through

cytoplasmic immune receptor tyrosine-activating domains that take on Src and Syk family kinases. In pro-Bells pro-BCR is potentially a functional receptor.

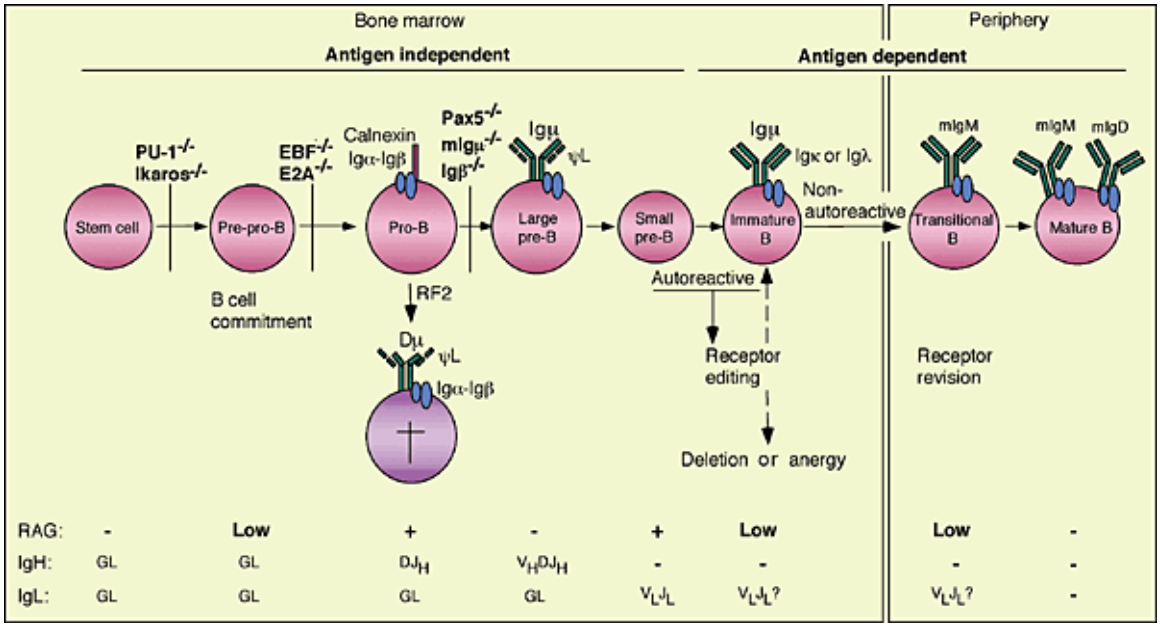


Figure 3. Development of B cells. RAG expression and rearrangements on both heavy (IgH) and light (IgL) chain genes are shown. GL is in germline configuration. Cell surface expression of pro-B (calnexin and Igα- Igβ), pre-B (Igμ, ψL and Igα- Igβ) or B specific (Igμ, Igκ or Igλ and Igα- Igβ) cell receptors is shown (6).

Immunoglobulin gene rearrangement starts with diversity (D) joining to junction (J_H) segment in pro-B-cells (Figure 3), and it is in the pro-B cells that antibodies first apply their regulatory function in B cell development. Because of random nucleotide loss and addition, D segments can be joined to J_H in any one of three reading frames, but there are only a few mature B cells in the mouse that have D segments in reading frame 2 (RF2), a phenomenon referred to as RF2 counter selection. DJ_H joins in RF2 encode a shortened form of mIgμ (Dμ) that associates with Igα - Igβ and surrogate light chains (ΨL, V-pre-B and λ5) to produce a defective pre-BCR that inhibits subsequent V_H to DJ_H

recombination and is unable to support further B cell differentiation. B cells that express Dμ are therefore arrested at the pro-B cell stage where they are either deleted or their truncated receptors are replaced by continuing recombination (Figure 3). Dμ signaling through Igα - Igβ is required for RF2 counter selection because in the absence of the transmembrane domain of mIgμ or Igβ there is no counter selection.

After DJ_H rearrangement, V_H genes become available to the V (D) J recombination and help complete the heavy chain transcription unit. The exchange from DJ_H to V (D) J_H recombination in pro-B cells is probably regulated at the level of V_H gene accessibility and appears to require Pax-5 and interleukin 7 (IL-7). B cells with complete V (D) J_H rearrangements also fail to accumulate in $Ig\beta^{-/-}$ mice. In each of these cases, impaired accumulation of V (D) J_H joins could result from a direct effect on recombination or alternatively could be due to a failure to positively select cells with complete receptors. V_H gene accessibility is dependent on transcriptional regulatory elements including the immunoglobulin heavy chain enhancer and is associated with the onset of germline transcription of V_H genes. It has been proposed that self transcription give V_H gene segments accessibility to the recombinase. An alternative possibility is that cis regulatory elements in immunoglobulin promoters and enhancers recruit factors that remodel chromatin domains and make V_H genes accessible for recombination independent of transcription.

The next step is to produce a pre-B cell that is expressing both low levels of surface and high levels of cytoplasmic μ heavy chains. Pre-B cells undergo light chain gene recombination. Successful light chain gene rearrangement leads to BCR assembly and replacement of the ΨLs in the pre-BCR by Igα or Igλ. Finally, the light- chain genes

are rearranged and the cell, an immature B cell, expresses both light chains (L chains) and μ heavy chains (H chains) as surface IgM molecule. Immature B cell differentiates within a few days into mature B cell and it has IgM and IgD (8).

Both B and T cells undergo positive and negative selection in the primary lymphoid organs. Positive selection requires signaling through the antigen receptor for the cell to survive. Developing B cells are positively selected when the pre-B receptor binds its ligand. (Developing T cells are positively selected for their ability to bind MHC as well as peptide.) Negative selection means that binding to the receptor results in cell death. Both immature B and T cells are negatively selected if they bind self antigen (8.9).

Development of T cells

T cell develops from bone marrow stem cells and their progenitors migrate to the thymus at a very early stage where they mature (5, 9). In the thymus the immature T cell undergoes several steps of receptor rearrangement and differentiation of progenitors. T cells have receptors specific for antigen. These T cell receptors (TCR) are generated by the rearrangement of germline genes. It occurs as the cells pass through the thymic cortex.

The TCR consists of two chains, alpha and beta, which is associated with the CD3 complex that signals to express CD4 and CD8. The TCR genes consist of a variable (V) gene segment, diversity (D) gene segment and a constant (C) gene segment, which are encoded by separate exons. In the TCR β chain, D to J rearrangements precedes V to DJ rearrangements. In the TCR α chain, gene rearrangements take place between V and J gene segments.

The enormous diversity of TCR molecules, needed to recognize a wide variety of antigens, is generated via the rearrangement of the gene sequences.

The Role of the pre-TCR During T cell Development

During early T cell development, transition from the $CD4^-8^-$ to the $CD4^+8^+$ stage has been identified. This occurrence is controlled by an immature form of the αβTCR, which has been termed the pre-αβTCR. This complex consists of a TCRβ chain, CD3 chains, and a newly identified protein which pairs with the TCRβ chain in the absence of the TCRα chain. This protein is called pre-Tα. The pre-TCR is essential since it allows cells bearing fully rearranged, in frame TCRβ chains, to be selected away from cells bearing non-functional TCRβ rearrangements. This selection therefore allows survival of potentially useful cells and destruction of useless cells. Although the selection mechanism is unclear, it is thought that successful pairing of a fully rearranged in frame TCRβ chain with the pre-Tα chain signals the thymocyte to proliferate, express CD4 and CD8 molecules, and initiate TCRα rearrangements. Thus, successful pre-TCR selection results in a large population of $CD4^+8^+$ thymocytes bearing in frame TCRβ chains, which can then be paired randomly with TCRα chains, to form the immature αβTCR collection.

Selection of the α and β T cell Receptor

The immature thymocytes are generated in the thymic cortex, and subsequently selected by the pre-TCR; the thymus then performs its second important function. Immature αTCR and βTCR bearing $CD4^+8^+$ thymocytes are subject to two selection processes, positive selection and negative selection. Positive selection, allows those cells that have the potential to recognise foreign peptides in association with self-MHC to mature to functional T cells and negative selection removes potentially dangerous T cells

that recognise self-peptides. Importantly, positive and negative selection processes can operate at the same time during development, so that a thymocyte can undergo negative selection without first being positively selected. T cells recognise antigen on presenting cells presented in the context of MHC molecules. It is important therefore that T cells are able to recognise peptides bound by self MHC antigens in order to function appropriately.

It is the cortical epithelial cells that impose positive selection on differentiating thymocytes. It appears that thymocytes which come into contact with and recognise MHC/peptide complexes on the epithelial cells are selected or permitted to continue their differentiation. Cells which fail to recognise MHC molecules on epithelial cells are not selected and die by a death-by-neglect mechanism. However, the process of negative selection, which induces into apoptosis those cells bearing potentially autoreactive TCRs, also involves recognition of peptide/MHC complexes in the thymus, although it is generally thought that this selection process is mediated by dendritic cells. Nevertheless, both positive and negative selection involve TCR binding to peptide/MHC complexes in the thymus, and it is still unclear how both events can operate on $CD4^{+}8^{+}$ thymocytes without cancelling each other out (9).

1.3 Transcriptional Regulation in Lymphocyte Development

It is well known that lymphocyte development originates from hematopoietic stem cells and goes through a series of growth and differentiation processes in the bone marrow and thymus. In these processes a number of transcription factors are involved in controlling the specific action and timing of gene expression. The E2A transcription factor is important during the initiation and progression of lymphocyte differentiation.

Accumulated evidence demonstrates that mammalian E2A transcription factor has a central role through the early B and T lymphocyte development (10).

For B cell, during development transcriptional regulation is aided by a complex set of transcription factors. During the differentiation of B cell, Ig gene rearrangements occur and at that time transcription factors play important roles (11). In the pre-pro B stage of the B cell XBP1 transcription factor is expressed which is important for exocrine gland and skeletal development (12). Another transcription factor, early B cell factor (EBF), is crucial for the development of B lymphocytes. This protein is expressed from the earliest stages of B cell development until the mature stage (13). STAT5 (Signal transducers and activators of transcription) plays key role in growth factor-mediated intracellular signal transduction in B cell (14). Ikaros transcription factor helps to regulate B lymphocyte differentiation (15). Researchers have found that Gfi1 gene expression is highest in early B cell subpopulation and is also differentially expressed during T cell development with peak levels at stages where pre-TCR expression or positive/negative selection take place. Gfi1 is absent in mature B cells (16). OBF-1 proximal promoter is crucial for activity in B cell (17).

During T cell development, transcriptional regulation is helped by a complex set of transcription factors. From T-lineage specification to peripheral T cell specialization the roles of transcription factors are different and important in gene regulation. For example, the functions of GATA-3, E2A/HEB, Id proteins, c-Myb, TCF-1, and members of the Runx, Ets, and Ikaros families are critical (18). There are several transcription factors which are expressed in different developmental stages of the T cell. In pro-stage of T-cell SCL- TAL (T cell acute lymphocyte leukemia 1), LMO1 (LIM domain only 1),

LMO2 (LIM domain only 2), E2A and HEB transcription factors are expressed (19). Transcription factor Notch1 is essential in T cell lineage commitment (20). PU.1 and GATA-3 (GATA binding protein 3) are transcription factors that are required for development of T cell progenitors from the earliest stages (21). The zinc finger transcription factor Th-POK regulates CD4 versus CD8 T cell lineage commitment (22).

1.2. Importance of the Subject

To start transcription for a particular gene, one or more transcription factors have to be bound to several specific binding sites. These binding sites are located in the regulatory region of the gene. A single transcription factor can be bound to multiple binding sites but they must have similar length and DNA sequence pattern. These binding sites are called as motifs. A motif is defined to be a short segment that occurs frequently in a DNA sequence. Since the majority of the motifs are unknown to us, our task is to find such motifs. The discovery of motifs will allow the biologist to understand the varied and complex mechanisms that regulate gene expression.

Currently computational tools are becoming increasingly useful in the explanation of various mechanisms in the process of finding of motifs, which has traditionally been studied by experimentation. Wet laboratory experimentation consists of studying one gene at a time although recent advances in microarray and chromatin immunoprecipitation (ChIP) technologies have made available large data sets for informatics analysis. Another impetus for computational approaches has been the explosion of annotated mammalian genomic data in various databases. Traditionally, DNA sequences upstream of the expressed genes (cis-acting) and transcription factor molecules binding to these DNA sequences (trans-acting) have been explored. We are

applying computational methods to identify transcriptional control elements in the DNA (promoters) of genes that may distinguish between the development of B and T cells.

1.3. Knowledge Gap

Understanding gene regulation within lymphocyte development is one of the major scientific challenges in the post-genome era. At the transcriptional level short segments of DNA occurring close to the start of a gene, known as TFBS or motifs, are believed to be involved in initiating the process of gene regulation.

Researchers are trying to find out which of the transcriptional control elements are active in B cell and T cell genes. They are also trying to find different transcriptional control programs active at different development stages of B cells and T cells.

Computational tools can help to find transcriptional control elements in lymphocyte development. These methods will take less time as compared to experiments performed in laboratory. Although not all of these analyses may identify biologically meaningful transcriptional control elements involved in lymphocyte development, we hope that they will provide a guide for the experimental biologist to focus on certain elements for biological validation.

2. BACKGROUND

2.1. Related Research in the Field of Detection of Transcriptional Control Elements

Identifying how genes are regulated is one of the great challenges of molecular biology. Major advances in this area are hoped for with the initiation of large scale gene expression profiling studies. Already many regulatory regions of genes and the binding sites of transcription factors have been biologically characterized. Databases now provide access to the weight matrices or consensus sequences that describe sites. Attempts are also being made to predict regulatory elements in non-coding genomic DNA through computational methods.

2.1.1 Experimental Methods

Chromatin Immunoprecipitation (ChIP) Experiments

Chromatin immunoprecipitation refers to a procedure used to determine whether a given protein binds to a specific DNA sequence in vivo. DNA-binding proteins are cross linked to DNA with formaldehyde in vivo. Chromatin is isolated and DNA bound to protein is sheared into small fragments. Then antibodies specific to the DNA-binding protein is used to isolate the complex by precipitation and the cross-linking is reversed to release the DNA and digest proteins. PCR was used to amplify specific DNA sequences to see if they had been precipitated with the antibody. It is a very slow process in finding transcription factor binding sites because it determines one site at a time. For example Tal1/ SCL transcription factor binding site has been identified using the chromatin immunoprecipitation method (23). The association of metal response element - binding transcription factor-1 (MTF-1) on the metallothionein-I promoter was examined using chromatin immunoprecipitation (ChIP) method. The results demonstrated that c-fos is

rapidly recruited along with MTF-1 to metallothionein-I promoter in response to zinc or cadmium (24).

The ChIP scanning strategy efficiently identified and localized primary glucocorticoid receptor (GR) target genes in vivo (25). Another use of the chromatin immunoprcitation (ChIP) method is the identification of MoKA, a Novel F-Box protein that modulates Kuppel-Like Transcription Factor 7 Activity (26). E2A target genes were identified by using a gene tagging-based chromatin immunoprecipitation system in B lymphocyte development (27).

Microarray Methods

PCR amplification along with microarray analysis of expressed genes can be employed to detect coordinately regulated genes. These methods were used to examine C3a and LPS mediated gene regulation in human mast cells (HMC). PCR amplification confirmed the microarray analysis of IL-1β up-regulation following C3a/LPS stimulation in HMC-1 cells. The microarray analysis generated more information in this study and thus more work will be needed to clarify the effects C3a stimulation has on mast cells (28).

Similar transcriptionally regulated genes can be analyzed by high throughput gene expression profiling with DNA microarray. Such transcriptionally regulated gene expressions play an important role in myocardial remodeling. Studies have been performed on cardiac muscle gene expression with DNA microarrays followed by a computational strategy to identify common promoter motifs that respond to insulin-like growth factor 1 (IGF-1) stimulation in cardiac muscle cells. This analysis showed that the Sp1 binding site is a likely target of IGF-1 action (29).

2.1.2. Computational Methods in Transcriptional Regulation

Most of the recent computational approaches have better performance over older methods. Motif finding algorithms are taking advantage of better bioinformatics approaches. The identification of regulatory motifs is important for the study of gene regulation.

Several programs have been developed that help to search for regulatory sequence motifs.

There are different computational methods that are applied to search genomic conserved motifs active in transcriptional regulation.

Hidden Markov Model

HMMER is an implementation of profile hidden Markov model (HMM) methods for sensitive database searches using multiple sequence alignments as queries. There are nine built-in functions available which help to query the sequences in different ways. Scientists have developed this approach to search genomic databases for conserved motifs present in the β-defensin family using HMMER,, in combination with the basic local alignment search tool, BLAST. This approach was first used to identify candidate second-exon coding regions, and later applied to finding associated first exons. These findings demonstrate an important proof-of-principle for a genome-wide search strategy to identify genes with conserved structural motifs (30).

The search engine MAPPER based on HMM is used for identification, visualization and selection of putative TFBSs occurring in the promoter or other regions of a gene from the human, mouse, fly, worm and yeast genomes. They built 1,079 models of TFBSs using experimentally determined sequence alignments of sites provided by the TRANSFAC and JASPAR databases and used them to scan sequences of these species.

In addition it allows the user to upload a sequence to query and to build a model by supplying a multiple sequence alignment of binding sites for a transcription factor of interest. Due to its extensive database of models, powerful search engine and flexible interface, MAPPER represents an effective resource for the large-scale computational analysis of transcriptional regulation. In several cases tested the method identified correctly experimentally characterized sites, with better specificity and sensitivity than other similar computational methods (31).

Motif Discovery Scan (MDscan) Method

MDscan searches for DNA sequence motifs. It is used to find motifs in entire genomes. MDscan adopts two strategies, word enumeration and position-specific weight matrix to retrieve motifs. This is faster than several established motif finding methods like BioProspector. These programs examine a group of sequences that may share common regulatory motifs and output a list of putative motifs as position-specific probability matrices, the individual sites used to construct the motifs and the location of each site on the input sequences (32). Researchers have reported that using this method 25 significant motifs active in amino acid starvation response have been predicted. The 25 motifs can be organized into 15 groups, 8 of which represent previously known TF motifs. *Saccharomyces cerevisiae* species was used in this research (33).

Logistic Regression Models

NF-κB is an immune gene in the human genome that is important to understand immune mechanisms and immune disease. By fitting logistic regression models to the promoters of 62 known NF-κB-regulated immune genes, patterns of transcription factor binding in the promoters of genes with known immune function have been identified.

These patterns were used to scan the promoters of additional genes to find matches to the patterns and selected those with NF-κB binding sites conserved in the mouse or fly and that are confirmed as NF-κB regulated immune genes based on expression data. From 6400 identified promoters in the human genome, only 28 predicted NF-κB target immune gene promoters, 19 of which regulate genes with known function (34).

oPOSSUM

The oPOSSUM system is used for identifying over-represented TFBSs in sets of co-expressed genes. This is based on web-based analysis of over-represented transcription factor binding sites. It combines a pre-computed database of conserved TFBSs in human and mouse promoters with statistical methods (35).

Multiple Expectation-Maximization for Motif Elicitation (MEME)

MEME is used for discovering motifs in a group of related DNA sequences. We have used this software in this project. It is discussed in more detail in Section 3 (Material and Methods).

MEME software is used to analyze consensus sequences for the ZAS family of proteins. The ZAS family is composed of proteins that regulate transcription by means of specific gene regulatory elements. Scientists proposed that the RSS are *cis*-acting DNA motifs which are essential for V (D) J recombination of antigen receptor genes. Due to its specific binding affinity for RSS and κB-like transcription enhancer motifs, they hypothesize that κB DNA binding and recognition component (KRC) may be involved in the regulation of V (D) J recombination. On the basis of that hypothesis they analyzed and obtained consensus sequence using MEME software (36).

Gibbs Recursive Sampler

Gibbs Recursive Sampler is an algorithm used to find motifs from DNA sequences. We have used this software in the project. Basic algorithm is discussed in Section 3(Material and Methods).

Heat shock (HS) genes have been identified in *Caenorhabditis elegans*. The upstream regions for these genes were analyzed using computational DNA pattern recognition methods. Two potential *cis*-regulatory motifs were identified. One of these motifs (TTCTAGAA) was the DNA binding motif for the heat shock factor (HSF), whereas the other (GGGTGTC) was previously unreported in the literature. Researchers determined the significance of these motifs for the HS genes using different statistical tests and parameters. Comparative sequence analysis of orthologous HS genes from *C. elegans* and *Caenorhabditis briggsae* indicated that the identified DNA regulatory motifs are conserved across related species. The role of the identified DNA sites in regulation of HS genes was tested by *in vitro* mutagenesis of a green fluorescent protein (GFP) reporter transgene driven by the *C.* elegans hsp-*16-2* promoter. DNA sites corresponding to both motifs are shown to play a significant role in the up-regulation of the *hsp-16-2* gene on HS. This is one of the extraordinary instances in which a novel regulatory element, identified using computational methods is shown to be biologically active. The contributions of individual sites toward induction of transcription on HS are non-additive, which indicates interaction and cross-talk between the sites, possibly through the transcription factors (TFs) binding to these sites (37).

2.2. Current Understanding of the Subject

Understanding how gene expression is regulated is one of the great challenges of molecular biology. It can be expected that this can be solved with the initiation of gene expression profiling studies**.** Already many regulatory regions of genes and the binding sites of transcription factors have been biologically characterized. Databases provide access to the weight matrices or consensus sequences that describe sites. Different datasets give easy ways to access transcriptional factors and provide information related to transcription factors. Several bioinformatics methods are useful to find transcription factor binding sites. Benefits of these methods are that it takes less time over traditional methods.

2.3. Research Objective

Our goal is to analyze and identify biologically meaningful transcriptional control elements involved in lymphocyte development. Using bioinformatics methods we hope to provide a guide for the experimental biologist to focus on certain elements for biological validation. This informatics approach to detecting transcriptional control elements may be a resourceful and successful help to biologists in the progress of study of transcriptional regulation that distinguishes B and T cell development.

3. MATERIALS AND METHODS

3.1. Data for the Project / Sample

We have employed three different datasets of sequences carrying potential transcriptional control elements active in regulating transcriptional events in lymphocyte development and one control dataset. The first dataset consisted of 356 upstream sequences of human immunoglobulin (Ig) and 242 upstream sequences of T cell receptor (TCR) genes (-1,000 to +10 with respect to TSS in both cases) collected from RefSeq (38).The second dataset was collected from the Eukaryotic Promoter Database (EPD) (39). From this database, promoter sequences were retrieved using keywords such as “B cell”, “T cell” and “Lymphoid”. The EPD dataset consisted of 62 upstream sequences from (-1000 to +50) including 24 specific for B cell terms, 30 specific for T cell terms and 15 for lymphoid. The third dataset is from microarray expression studies of up-regulated genes in B and T cells. In the case of B and T cells, there were 24 and 32 sequences (–1000 to +50 with respect to TSS) respectively from the genes showing up-regulated mRNA expression in the corresponding cell type (40, 41). The final dataset is a collection of human promoter sequences from Richard Myers’ Lab, which has been biologically validated (42).

These datasets are labeled as:

1) Ig, TCR (from REFSEQ)

2) EPD_B, EPD_T (From EPD)

3) Microarray_B, Microarray_T (From microarray expression studies)

4) Myers

3.2. Software

In this project we utilized several software and web-based tools to analyze the different datasets. EZ-Retrieve is a web-based tool for retrieving TFBS. Transcription Element Search System (TESS) is also a web-enabled tool to get consensus sequences of transcription factors. Multiple Expectation-Maximization for Motif Elicitation (MEME) and Gibbs Motif Sampler software were used to get potential regulatory motifs from upstream sequence of B and T cell-specific genes. Along with these tools, we also used various other software to analyze gene sequences.

EZ-Retrieve

The EZ-Retrieve tool is designed for retrieving any particular region of human genome sequence from the NCBI database and analyzes retrieved sequences for TFBS as they appear on the TRANSFAC database (43). This tool is web based, user friendly and used for sequence retrieval. EZ-Retrieve is available at the following web address: http://www.cag.icph.org/bioinformatics.html.

Transcription Element Search System (TESS)

TESS is a web tool for predicting TFBS in DNA sequences (44). It can identify binding sites using site names or consensus strings and positional weight matrices from the TRANSFAC, IMD, and CBIL-Gibbs Mat database. In this project we used the accession number of the TFBS to retrieve the consensus sequences. The articles referring the corresponding TFBS can also be found using TESS. We have also used the TFBS to find the information related to the target genes and their corresponding references from this website.

Multiple Expectation-Maximization for Motif Elicitation (MEME)

MEME is used for discovering motifs in a group of related DNA or protein sequences. MEME is based on the Expectation-Maximization (EM) approach. EM approach consists of two steps, expectation and maximization, which are repeated consecutively. In expectation step, the column-by column composition of the site is used to estimate the probability of finding motif at any position in each of the sequence. The maximization step uses the estimated probability of motif from the expectation step and multiplies that with the background frequency of the remaining positions. At this point the software calculates the likelihood, which tells us where motif matches positions A, B or others in the sequence 1. Similarly for other sequences the process is repeated. The expectation and maximization steps are repeated until the product of probability of motif and background is not constant (45).

MEME displays the occurrences (sites) of the motif in the training set. Each site is identified by the name of the sequence where it occurs, the strand, and the position in the sequence where the site begins. The occurrences of the motif in the training set sequences are shown with block diagrams. One diagram is printed for each sequence showing all the occurrences of the motif in that sequence. The sequences are sorted by the lowest p-value among all occurrences of the motif in a given sequence. The position-specific scoring matrix corresponding to the motif is printed for use by database search programs such as MAST. The motif itself is a position-specific probability matrix giving, for each position in the pattern, the probabilities of each possible letter occurring there. The probability

matrix is printed such that columns correspond to the letters in the alphabet and rows correspond to the positions of the motif.

In our project in order to use the MEME software, we employed the SSH software to connect to the AVIDD cluster on the IUPUI supercomputer (SP). We run the MEME software on the AVIDD clusters. We transferred the Fasta files of DNA sequences from personal computer to SP using SSH. Finally using command line we were able to run MEME on the various data sets. We performed MEME on the data sets namely, EPD_B, EPD_T and Microarray_B and Microarray_T, input in the Fasta file format.

Command which we have used to run MEME was

memejob -maxsize 60000 –dna p4 –nmotifs 5 –minw 6 –maxw 50 Filename

In the command, maxsize indicates the maximum size of file which is 60,000 nucleotides, nmotifs indicates the number of motifs, minw indicates minimum width of motif sequence (6) and maxw indicates maximum width of motif (50), and at the end the filename is written.

In result output was obtained as five different motifs with gene sequences, along with information content.

Gibbs Recursive Sampler

The Gibbs Recursive Sampler (46) is a software package for locating common elements in collections of biopolymer sequences. This software allows us to identify motifs from DNA sequences. This software is different from the EM algorithm; it is based on the Monte Carlo sampling algorithm. There are two steps which iterate many times. In the first step, a random motif is chosen from one sequence. In the second step,

that sequence is aligned back and forth until the motif probability to the background probability is maximized in each left out sequence. This will possibly provide a motif in each of the left out sequences. We will chose a new location for another motif and then repeat steps one and two.

We run the Gibbs Recursive Sampler in UNIX in order to obtain the result. We entered the command as,

–Pbernoulli Promoter_EPD.txt 50 5 –n –o promoterEPD.dat

where Promoter_EPD.txt is the input file in Fasta format, 50 indicates the size of motif, 5 indicates number of motifs, -n indicates the nucleotide sequence, -o indicates the output sequence and promoterEPD.dat indicates the name of output file.

We have performed Gibbs Recursive Sampler process for data sets, namely, EPD_B**,** EPD_T and Microarray_B and Microarray_T. In output, we obtained one motif for each dataset and for each motif a description of the start of sequence, end of sequence, the sequence, and the probability of the motif is shown.

ClustalW

This software is used for multiple sequence alignment of DNA or protein sequences (47). It is used to calculate the alignment score of sequences. In this project we used ClustalW to calculate the alignment score of the motifs, which we retrieved from MEME and Gibbs Recursive Sampler software.

3.3. Procedure

The first step of the project is to isolate upstream sequences of human and mouse genes involved in B cell and T cell development.

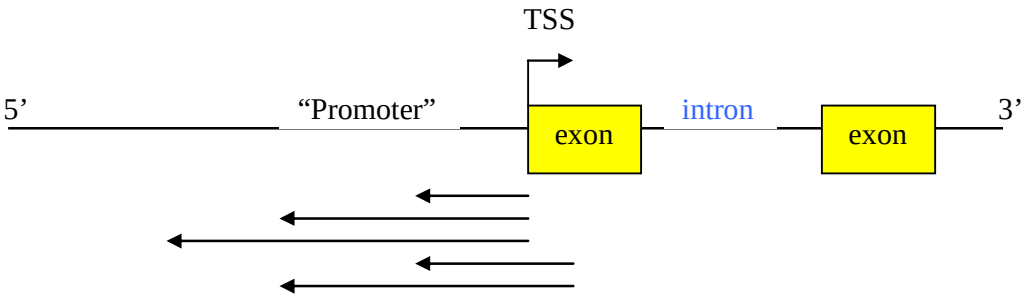


Figure 4. Upstream Sequences of a Gene for Analysis

We have employed four different datasets to identify TFBS specific to lymphocyte development.

3.3.1 Gibbs Sampler Identification of Motifs in Ig and TCR Genes

Gibbs Recursive Sampler was performed on Ig and TCR genes obtained from RefSeq. One predominant sequence was observed from each of the Ig and TCR genes.

3.3.2 Identification of TFBS Enriched in Ig and TCR Genes

We performed EZ-Retrieve on the first dataset, which was collected from RefSeq (NCBI) on both Ig and TCR genes. The results obtained from EZ-Retrieve were the different TFBS. We then performed EZ-Retrieve on the control Myers dataset, and obtained the TFBS. Comparisons of the TFBS of Ig and TCR genes of RefSeq dataset with the TFBS of Myers’ dataset were made. We labeled the TFBS which were enriched in TCR genes compared to the Myers’ dataset as Tm and these enriched in Ig genes compared to the Myers’ dataset as Im. Similarly, we labeled the TFBS, which were enriched in TCR genes, but not in Ig genes as B; TFBS that are enriched in Ig genes, but

not in TCR genes are called as A. When we combine the various lists of TFBS as Im U A and Tm U B, we obtained a list of B specific TFBS and a list of T specific TFBS, respectively as shown in Figure 5. This methodology is shown in the flowchart in Figure 6.

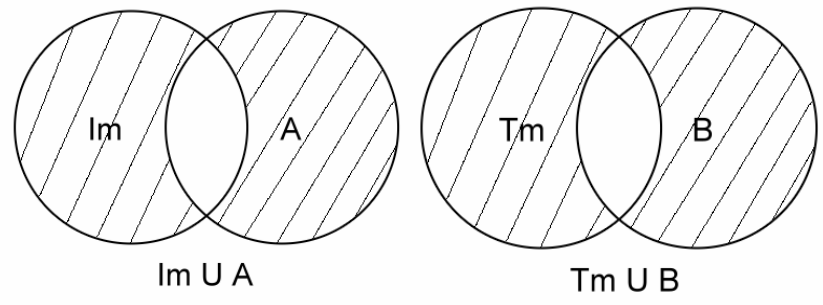


Figure 5. Venn diagrams for B and T cell specific TFBS. Im U A represents B specific TFBS and Tm U B represent T specific TFBS.

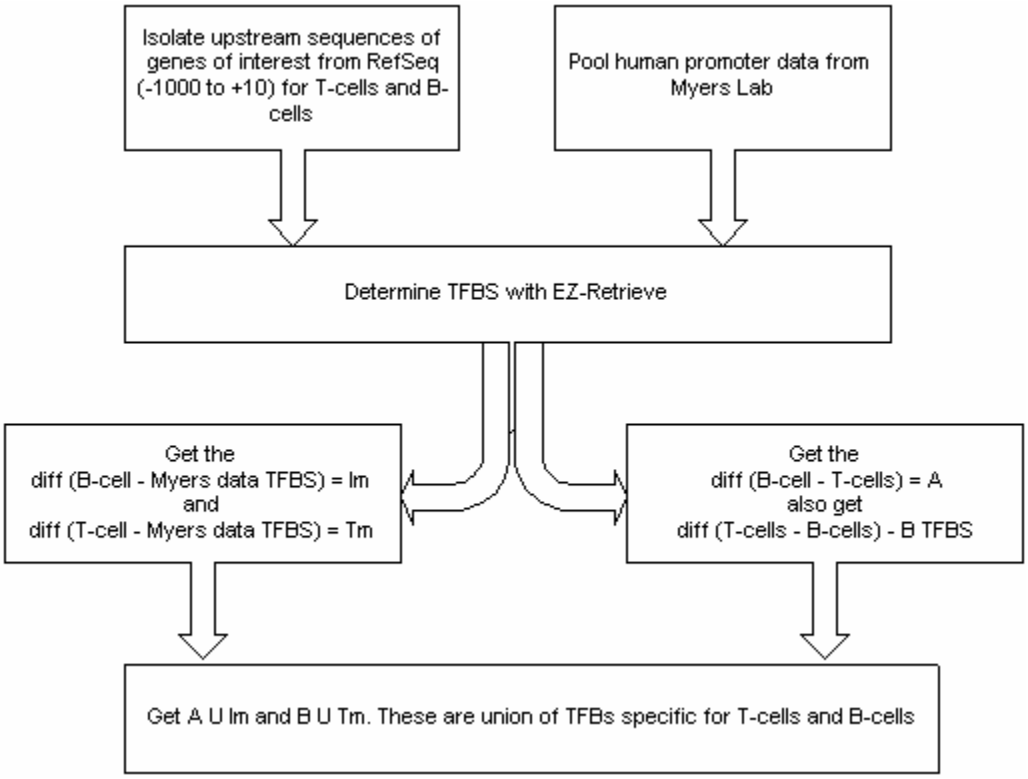


Figure 6. Flowchart of methodology for the identification of B and T cell specific TFBS. Im, indicates the list of TFBS which are present in Ig but not in the Myers dataset. Tm indicates the list of TFBS which are present in TCR but not in the Myers dataset. Similarly, A indicates list of TFBS which are found in Ig but not in TCR. B indicates the list of TFBS which are found in TCR but not found in Ig.

3.3.3 Motif Discovery Using MEME and Gibbs Sampler on the EPD and Microarray Datasets

We performed MEME on EPD_B and EPB_T of the Eukaryotic Promoter Database (EPD) datasets and Microarray_B and Microarray_T of microarray datasets separately. We performed MEME on positive strand and negative strand. We arbitrarily chose a MEME parameter of five motifs. The motif sequences returned from MEME were then used as input to the EZ-Retrieve software. This gave the corresponding TFBS present in each motif sequence, which was retrieved from MEME for both datasets.

Similarly, we performed Gibbs Recursive Sampler on both datasets namely, EPD_B and EPB_T of Eukaryotic Promoter Database (EPD) and Microarray_B and Microarray_T of the microarray dataset, which resulted in one motif sequence for each dataset. We proceeded to use this motif for EZ-Retrieve to get TFBS. We performed Gibbs Recursive Sampler for positive and negative strands. We compared common motifs identified by the Gibbs Recursive Sampler and MEME software. The methodology for this is shown in the flowchart in Figure 7. We calculated alignment score of the motifs, which were obtained from MEME and Gibbs Recursive Sampler using ClustalW.

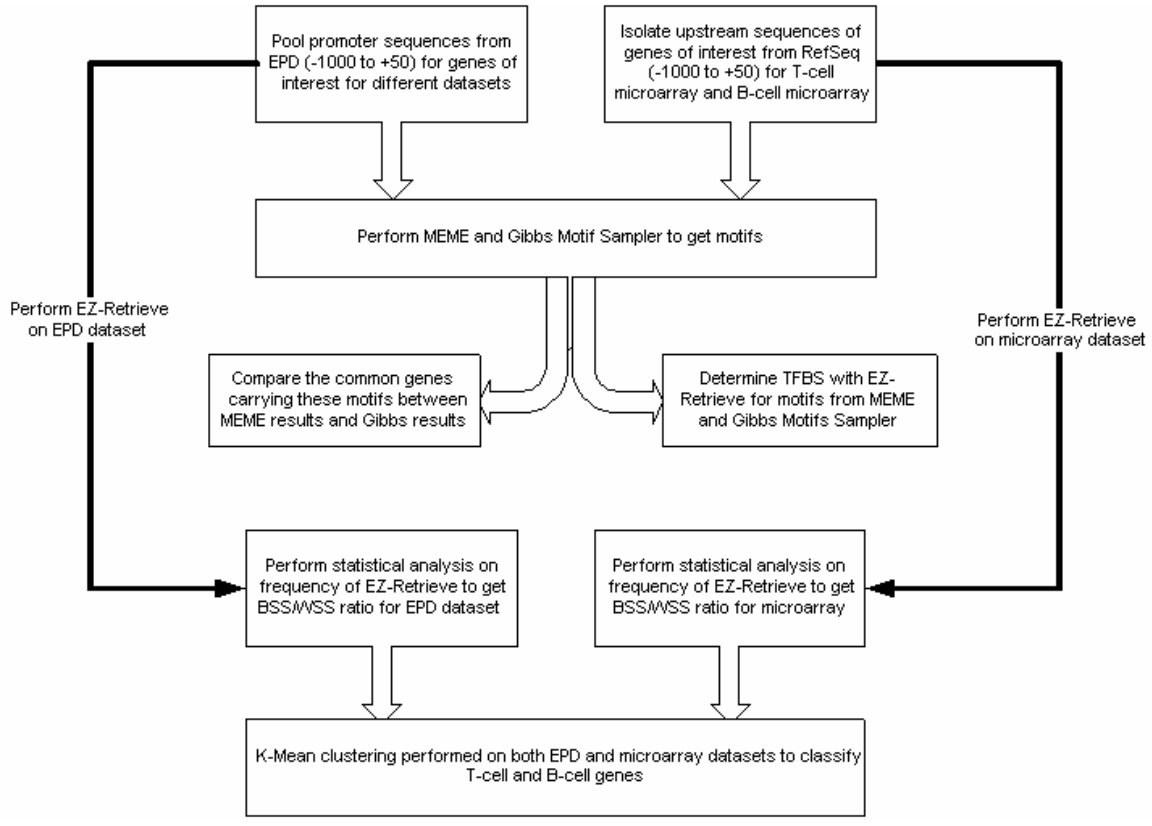


Figure 7. Flowchart showing the analysis of the EPD and microarray data

3.3.4 Statistical Method (BSS/WSS)

This is a purely statistical method (48) which is used to ascertain the developmental distinction between B and T cell genes based on normalized frequencies of TFBS which are obtained from EZ-Retrieve for the EPD_B and EPD_T as well as Microarray_B and Microarray_T datasets. Analysis is then performed on the basis of Between Sum of Squares (BSS) and Within Sum of Squares (WSS) methods across the two groups (B cell and T cell). For these normalized frequencies of TFBS, BSS and WSS were calculated across the two groups based on the equations given below.

$$BSS = (\overline{X} - a)^2 n_i + (\overline{Y} - a)^2 n_j$$

$$WSS = \{(X_1 - \overline{x})^2 + (X_2 - \overline{x})^2 + ... + (X_n - \overline{x})^2\} + \{(Y_1 - \overline{y})^2 + (Y_2 - \overline{y})^2 + ... + (Y_n - \overline{y})^2\}$$

$$\overline{x} = \frac{(x_1 + x_2 + x_3 + \cdots + x_i)}{n_i} \quad , \quad \overline{y} = \frac{(y_1 + y_2 + y_3 + \cdots + y_i)}{n_j}$$

where x_1 and x_2 … x_i is the normalized frequency of one group; y_1 and y_2 … yj is the normalized frequency of the other group.

$\overline{x}$ and $\overline{y}$ are the average normalized frequency for each group and n_i and n_j are number of genes in each groups (B cell and T cell respectively).

$$\overline{a} = \frac{(\overline{x} + \overline{y})}{2}$$

where ‘a’ is the sum of these two averages divided by the number of groups, that is it is the average of averages. The number of groups in our case is 2, corresponding to B and T cells.

Once BSS and WSS are calculated the BSS/WSS ratio of TFBS among the two groups is easily found. Similarly, for the microarray genes BSS/WSS ratio can be calculated. For bioinformatics study, we have used the TFBS with a BSS/WSS ratio above the cut-off value of 0.05 for the EPD genes and above 0.54 for the microarray genes. In the case of the EPD group, there were nine TFBS above the cut-off value. For the microarray dataset, there were ten TFBS above the cut-off value.

K-Means Clustering

One of the most commonly used clustering algorithms is the k-means clustering technique (49). The idea behind this is to identify a group of patterns from within the entire data set, which are sufficiently ‘close’ or ‘similar’ to each other. Closeness is usually measured by some sort of distance; the most commonly used being the Euclidean distance.

K-means clustering was applied on the results which we obtained from statistical analysis of the normalized frequencies of TFBS. The data consisted of B and T cell genes in each group. There were a total of nine TFBSs in the EPD group and ten TFBSs in the microarray group that were employed to classify the clusters.

EPCLUST is web-based software which allows a user to enter their data file and select various criteria, such as the type of clustering to perform and number of clusters to output. In our case the number of clusters was two, one for B cell and one for T cell, so we could classify the groups into two clusters.

4. RESULTS

In this section, the results from various prediction and analysis methods related to transcriptional control elements in lymphocyte (B and T cells) are presented.

Gibbs Sampler on Ig and TCR Datasets

Two predominant pattern sequences were obtained from the 356 Ig sequences and 242 TCR sequences using Gibbs Recursive Sampler as shown in Table 1. A biologically validated TF binding site Oct-1 is found within the motif identified from the Ig genes (shown as the sequence in bold in Table 1).

Lymphocyte	Pattern
B-cell	(C/T)**ATGCAAAT**(C/A/G/A)
T-cell	(A/G)GTGACATCA

Table 1. Pattern search on B and T cell-specific genes

EZ-Retrieve on B cell and T cell Genes

We performed EZ-Retrieve (43) as discussed in Section III on the Ig and TCR genes datasets and obtained a list of TFBS as B cell specific and T cell specific after comparing them with the EZ- Retrieve results of the Myers dataset. This list implies that the TFBS enriched in TCR but not in the Ig and vice versa. TFBS with low frequencies are found in Ig but not in TCR and vice versa. For example B cell specific GATA-1 is not found in TCR specific genes and TCR specific c-Ets is not found in Ig specific genes. Higher probability TFBS are found in both TCR and Ig. These results are shown in Table 2.

B-cell Specific		T-cell Specific	
diff(B - T)		diff(B - T)	
Accession No.	**TFBs**	**Accession No.**	**TFBs**
M00075	GATA-1	M00041	CRE-BP
M00161	Oct-1	M000209	NF-Y
M00227	v-Myb	M00074	c-Ets-
M00117	C/EBPb	M00074	HFH-1
M00086	Ik-1	M00208	NF-kap
M00051	NF-kap	M00249	CHOP-C
M00221	SREBP-	M00156	RORalp
M00246	Egr-2	M00050	E2F
M00245	Egr-3	M00146	HSF1
M00210	OCT-x		
M00033	p300		
M00243	Egr-1		
M00248	Oct-1		
M00059	YY1		
M00070	Tal-1b		

Table 2. Comparison of TFBS of B and T cell-specific genes

Common Genes from MEME and Gibbs Recursive Sampler

We have performed MEME and Gibbs Recursive Sampler on the different datasets and find common genes carrying different motifs. Common genes carrying various motifs predicted by MEME and Gibbs Recursive Sampler on the different datasets have been tabulated in the Table 3. A number of these genes have been implicated in lymphocyte development.

From the Microarray_T dataset, MEME and Gibbs Recursive Sampler identified one motif as the same (MEME motif 1) where as from the Microarray_B dataset both algorithms identified one motif (MEME motif 2) that was same in the reverse direction (one motif in + strand compared to second in – strand).

	Common Genes in the different datasets	
Microarray_B		
Gibbs	**MEME**	**Gene Name**
X13450	X13450	Cd79a - B- cell antigen receptor complex
M84756	M84756	Ank1 - Ankyrin 1
AK010422	AK010422	Hba-a1 - Hemoglobin alpha, adult chain
BC063060	BC063060	1Tal1 - T-cell acute lymphocytic leukemia
BC006967	BC006967	Dusp1 - Dual specificity protein phosphatase 1
Microarray_T		
Gibbs	**MEME**	
X70991	X70991	NAB2 - NGFI-A binding protein 2
BC022522	BC022522	CD200 antigen
BC006196	BC006196	TNFRSF9 - tumor necrosis factor receptor superfamily, member 9
AF064090	AF064090	TNFSF14 - tumor necrosis factor (ligand) superfamily, member 14
EPD_B		
Gibbs	**MEME**	
EP26038	EP26038	Hs IL-4 (BSF-1)
EP16056	EP16056	Hs renin
EP74540	EP74540	Hs CD79A
EPD_T		
Gibbs	**MEME**	
EP11142	EP11142	Hs IL-2 receptor P1
EP11143	EP11143	Hs IL-2 receptor P2

Table 3. Comparison of common genes carrying motifs identified by MEME and Gibbs recursive sampler

The percentage of genes carrying various motifs and the information content of the MEME motifs for EPD genes are given in Table 4. The information content diagram provides an idea on which positions in the motif are most highly conserved and it gives a

measure of the usefulness of the motif for database searches. Information content is equal to the log likelihood ratio divided by the number of occurrences times ln (2).

B cell genes

	Sites	Percentage	Information Content
Motif 1	12	50	41.3
Motif 2	20	83.3	17.3
Motif 3	9	37.5	46.1
Motif 4	7	29.1	41.7
Motif 5	7	29.1	43.0

T cell genes

	Sites	Percentage	Information Content
Motif 1	12	40	35.0
Motif 2	13	43	31.4
Motif 3	7	23	50.7
Motif 4	5	16.6	65.5
Motif 5	7	23	46.5

Table 4. Proportion of conserved motifs in EPD target genes.

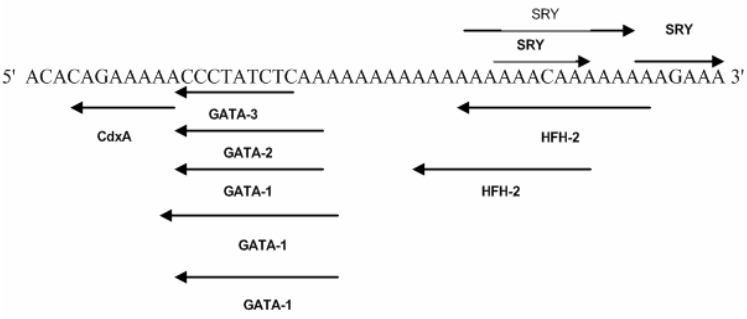
EZ-Retrieve Performed on MEME Motifs

We performed EZ Retrieve on all MEME motifs from the different datasets in order to find TFBS and their location and direction on the motif. Figure 8 shows the TFBS predicted by EZ-Retrieve in various MEME motifs. For example, Figure 8A shows Motif -2 from the Microarray_B dataset in which eleven TFBS are predicted. Three SRY sites are in the positive direction. Two HFH-2, three GATA-1, one GATA-2, one GATA-3, one CdxA sites are in the negative strand.

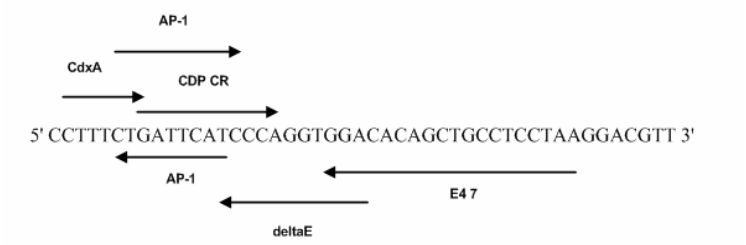
Similarly, Figure 8B shows motif -3 from EPD_T in which six TFBS are predicted. AP-1, CDP CR, CdxA TFBS are in the positive direction and E47, AP-1 and deltaE are in the negative stand.

Figure 8C shows motif -3 from Microarray_B in which six TFBS are predicted. One HFH-2, one SRY TFBS are in the positive direction and two SRY, one TATA, and one Pbx-1 are in the negative strand.

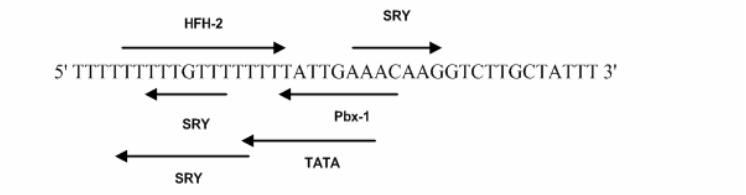
Finally, Figure 8D shows motif-4 EPD_B in which four TFBS are predicted. One Gfi-1, one AML-1a, one Oct-1 are in the positive direction and one CdxA site is in the negative strand.



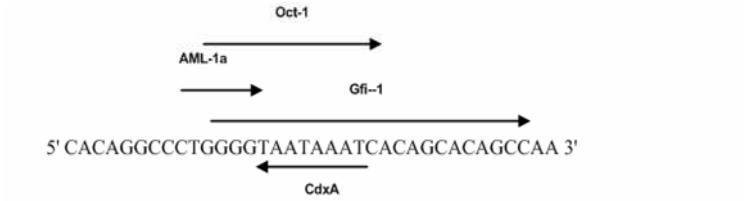
A. Microarray_B Motif 2



B. EPD_T Motif 3



C. Microarray_B Motif 3



D.EPD_B Motif 4

Figure 8. Examples of motifs analyzed by EZ-Retrieve. The sites in the positive strand are shown above and ones in the negative strand below the motif sequence. Some of these TFBS have been implicated as sites for binding immune-specific TFs (*e.g.*., GATA, deltaE, E47, Oct-1) and maybe part of potential biologically meaningful TCEs.

Alignment Score

In order to look for identity between sequences in the different motifs obtained from the MEME and Gibbs Recursive Sampler software we used ClustalW for alignments as discussed in Section III. The alignment scores between the Gibbs Recursive Sampler motif and the different MEME motifs are higher compared to those obtained between MEME motifs aligned with each other. This indicates that MEME motifs are distinct patterns as identified by the software.

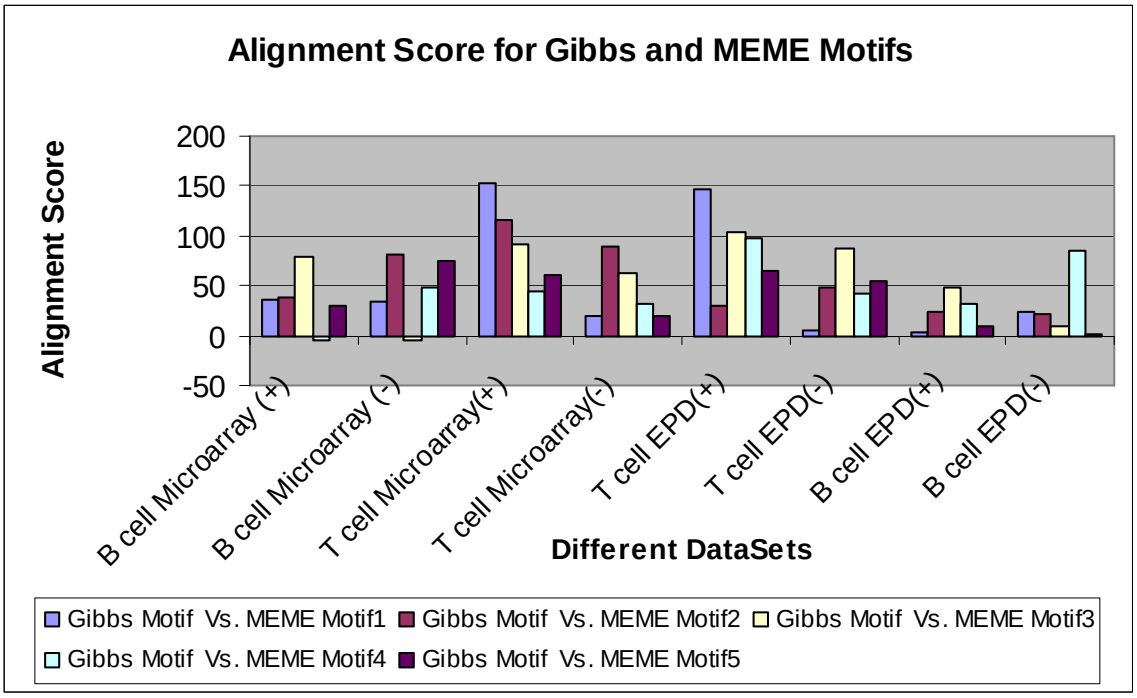


Figure 9. Pair-wise alignment between various motifs. ClustalW was employed to align pairs of motif sequences identified by MEME with their corresponding one identified by Gibbs. '+' indicates positive strand and '-' indicates negative strand sequences (orientation)

When the different MEME motifs were aligned with each other, we did not notice any significant sequence homology. When we compared alignment scores between motifs from T cell EPD (+), we find Gibbs Motif 1 and MEME Motif 1 alignment score

to be high (150). The alignment scores for MEME and Gibbs Recursive Sampler motifs are shown in Figure 9.

4.1 Statistical Analysis to Distinguish B and T cell Genes

As discussed in Section III, we followed a statistical approach to classify B and T cell genes based on their TFBS frequencies. We made a discriminate prediction, which gave support to our hypothesis. We selected TFBS based on a threshold BSS/WSS ratio of TFBS frequencies that could discriminate between the two groups (B cells and T cells), for both the EPD and microarray datasets. We had determined BSS and WSS by using a statistical approach. After finding BSS and WSS we evaluated the BSS/WSS ratio for both datasets. Then, we plotted a graph of BSS/WSS *Vs* TFBS and identified a cut-off value from this graph. For the EPD dataset, such a graph is shown in Figure 10 and the cut-off value is 0.05. For the microarray dataset, the graph of BSS/WSS *Vs* TFBS to find the cut-off value is shown in Figure 11 and cut-off value is 0.054.

The information related to the TFBS found above the cut-off value for EPD dataset and microarray dataset are shown in Tables 5 and 6 respectively. In case of the EPD dataset we have nine TFBS above the cut-off and in case of the microarray dataset we have ten TFBS. We have listed the description, BSS/WSS ratio, references and target gene of the corresponding TFBS.

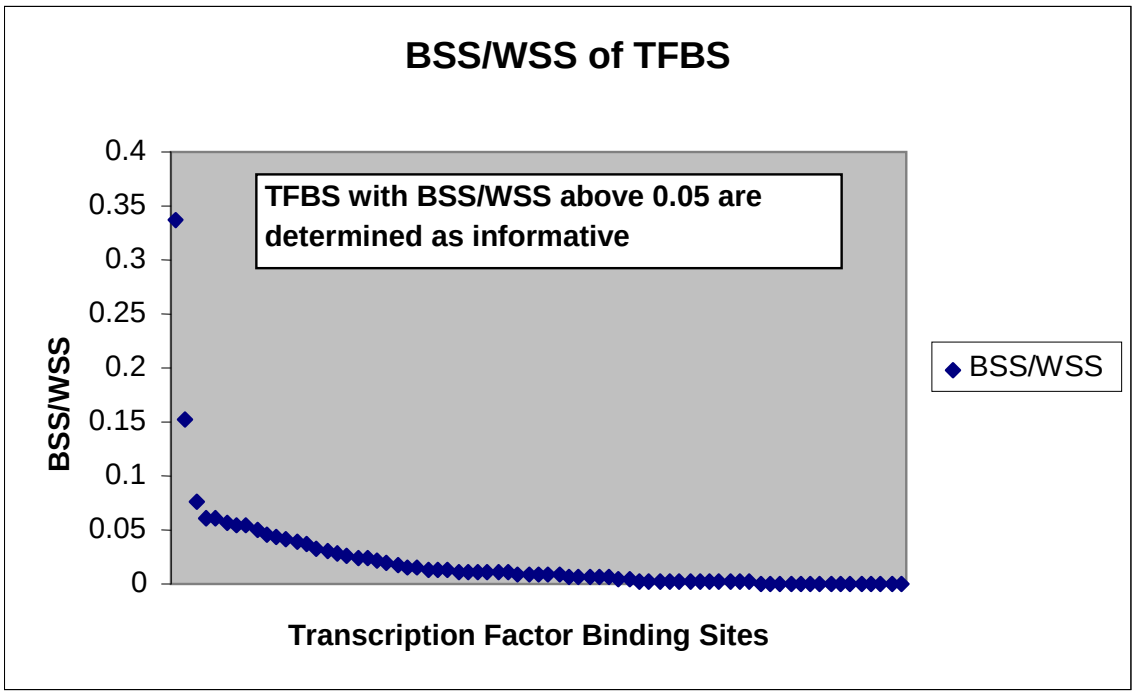


Figure 10. BSS/WSS ratio of TFBS in the upstream regulatory sequences of EPD genes

Transcription Factor Site	Description	BSS/WSS	References	Target Gene
MZF1	MZF1	0.336610435	Morris J. F., Hromas R., Rauscher III F. J., Mol. Cell. Biol. 14:1786-1795 (1994).	Zinc finger protein
Evi-1	ectopic viral integration site 1 encoded factor	0.152836062	Delwel R., Funabiki T., Kreider B. L., Morishita K., Ihle J. N., Mol. Cell. Biol. 13:4291-4300 (1993). Mol Cell Biol. 1991 May;11(5):2665-74. Perkins AS, Fishel R, Jenkins NA, Copeland NG.	Evi-1 protein
p300	p300	0.075223012	Rikitake Y., Moran E., Mol. Cell. Biol. 12:2826-2836 (1992). Rasti M, Grand RJ, Mymryk JS, Gallimore PH, Turnell AS.	AdE1A
CREB	cAMP-responsive element binding protein	0.060519545	Benbrook D. M., Jones N. C., Nucleic Acids Res. 22:1463-1469 (1994). Proc Natl Acad Sci U S A. 1992 Aug 1;89(15):7070-4 Zhao LJ, Giam CZ.	HTLV-I 21-base-pair repeats
HSF2	heat shock factor 2	0.060234668	Kroeger P. E., Morimoto R. I., Mol. Cell. Biol. 14:7592-7603 (1994). Sistonen L, Sarge KD, Phillips B, Abravaya K, Morimoto RI.	HSE
v-ErbA	viral homolog of thyroid hormone receptor alpha1	0.055842813	Subauste J. S., Koenig R. J., J. Biol. Chem. 270:7957-7962 (1995)	
SRY	sex-determining region Y gene product	0.054697372	Pontiggia A., Rimini R., Harley V. R., Goodfellow P. N., Lovell-Badge R., Bianchi M. E., EMBO J. 13:6115-6124 (1994). Proc Natl Acad Sci U S A. 2003 Jun 10;100(12):7045-50. Epub 2003 May 22. Harley VR, Layfield S, Mitchell CL, Forwood JK, John AP, Briggs LJ, McDowall SG, Jans DA.	
Sp1	Stimulating protein 1	0.053412382	Gomez et al1998 Ghadially et al2005	BCL2 CCL22
HNF-3b	Hepatocyte Nuclear Factor 3beta	0.050798278	2002 Dec;38(3):229-34., Seike M, Gemma A, Hosoya Y, Hosomi Y, Okano T, Kurimoto F, Uematsu K, Takenaka K, Yoshimura A, Shibuya M, Ui-Tei K, Kudoh S	BUBR1

Table 5. List of TFBS of EPD genes above the BSS/WSS threshold

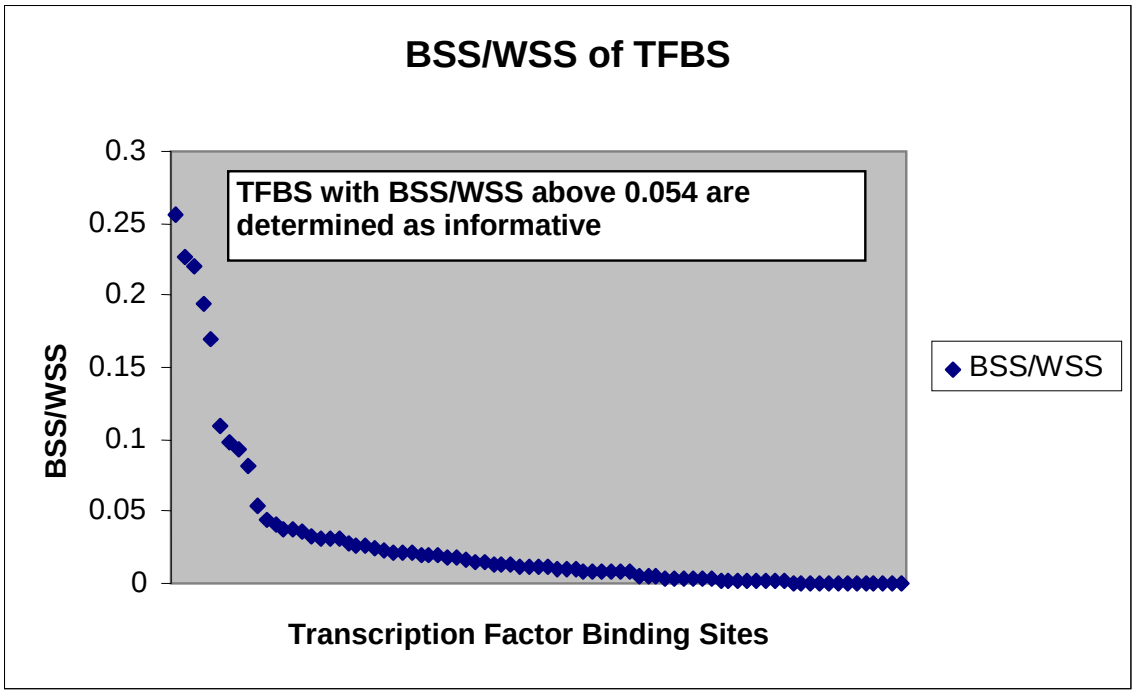


Figure 11. BSS/WSS ratio of TFBS in the upstream regulatory sequences of Microarray genes

Transcription Factor Binding Site	Description	BSS/WSS	References	Target Gene
Nkx-2.		0.255587454	2002 Dec;38(3):229-34., Seike M, Gemma A, Hosoya Y, Hosomi Y, Okano T, Kurimoto F, Uematsu K, Takenaka K, Yoshimura A, Shibuya M, Ui-Tei K, Kudoh S.	BUBR1
GATA-1	GATA-binding factor 1	0.226042137	MEDLINE:93309433, Merika M., Orkin S. H., Mol. Cell. Biol. 13:3999-4010 (1993). J Biol Chem. 2003 Nov 14;278(46):45620-8. Epub 2003 Aug 26. Ghirlando R, Trainor CD.	GATA-1
C/EBP	CCAAT/enhancer binding protein	0.220181319	MEDLINE:91187609, Grange T., Roux J., Rigaud G., Pictet R., Nucleic Acids Res. 19:131-139 (1991). Nucleic Acids Res. 1991 Jan 11;19(1):131-9 Grange T, Roux J, Rigaud G, Pictet R.	GRU
AML-1a	runt-factor AML-1	0.194237573	Meyers S., Downing J. R., Hiebert S. W., Mol. Cell. Biol. 13:6336-6345 (1993). Mol Cell Biol. 1993 Oct;13(10):6336-45. Meyers S, Downing JR, Hiebert SW.	AML-1/ETO
CdxA	CdxA	0.170092927	MEDLINE:94232818 Margalit Y., Yarus S., Shapira E., Gruenbaum Y., Fainsod A. Nucleic Acids Res. 21:4915-4922 (1993). Nucleic Acids Res. 1993 Oct 25;21(21):4915-22. Margalit Y, Yarus S, Shapira E, Gruenbaum Y, Fainsod A.	CdxA
GATA-2	GATA-binding factor 2	0.110005343	Merika M., Orkin S. H., Mol. Cell. Biol. 13:3999-4010 (1993).	GATA-2
GATA-3	GATA-binding factor 3	0.097416815	Glimcher et al2000	GATA-3
N-Myc	N-Myc	0.092165944	Alex R., Soezeri O., Meyer S., Dildrop R., Nucleic Acids Res. 20:2257-2263 (1992). J Biol Chem. 1994 Jan 21;269(3):1785-93. Draeger LJ, Mullen GP.	Helix-1
IRF-1	interferon regulatory factor 1	0.08167044	Tanaka N., Kawakami T., Taniguchi T., Mol. Cell. Biol. 13:4531-4538 (1993)., TRANSFAC Reports 1:0001 (1998).	IL12
Egr-2	Egr-2/Krox-20 early growth response gene product	0.054565406	Swirnoff A. H., Milbrandt J., Mol. Cell. Biol. 15:2275-2287 (1995). J Biol Chem. 1998 Oct 9;273(41):26923-30 Decker EL, Skerka C, Zipfel PF.	IL-2

Table 6. List of TFBS of Microarray genes above the BSS/WSS threshold

We have performed k-means clustering on the EPD and microarray datasets in order to classify B cell and T cell genes in different clusters using TFBS frequencies. We set two as the input parameter for the number of clusters in order to verify the classification of two sets of genes – one B cell specific and the other T cell specific. In the first cluster there were forty-one genes and in the second cluster there were thirteen genes. Figures12 and 13 show the first and second clusters of EPD dataset genes, respectively. In case of the microarray dataset, the first cluster showed fifty- six genes and the second cluster only one gene. These are represented in Figures 14 15. The clustering results, which we obtained, do not classify B and T cell groups efficiently. The reason could be due to the poor discriminating power of the TFBS frequencies in the B and T cell gene datasets to distinguish between groups (*ie* low BSS/WSS values).

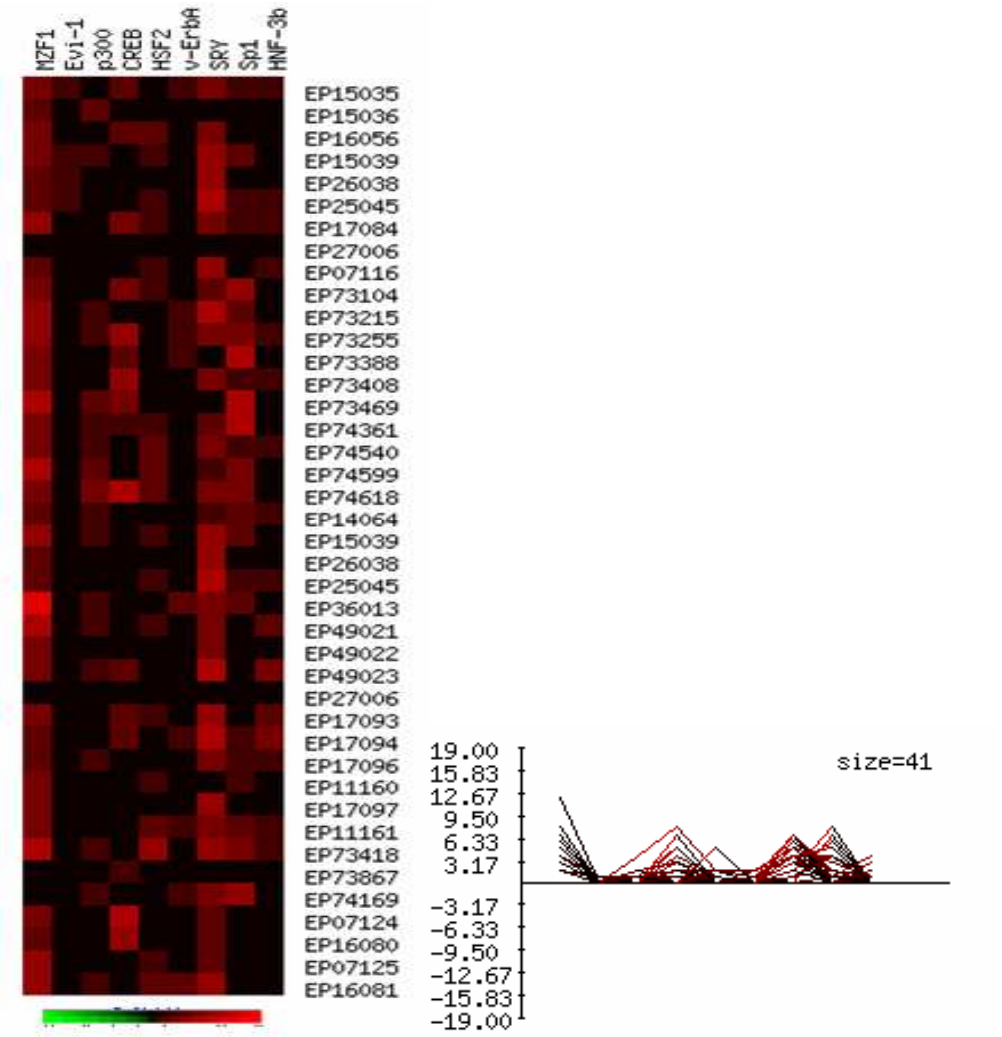


Figure 12. Cluster 1 from EPD genes

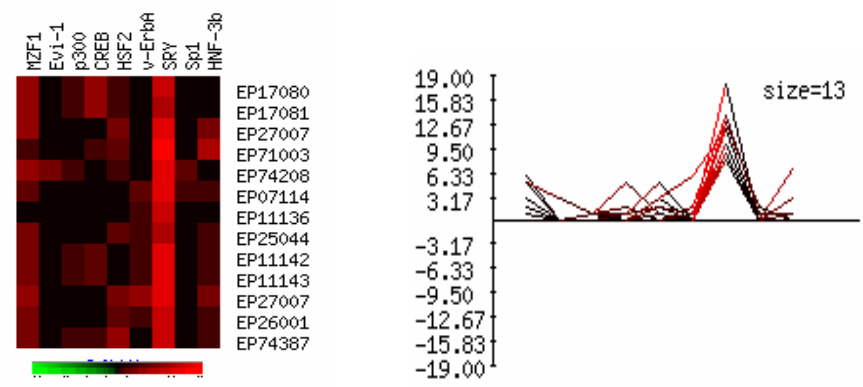


Figure 13. Cluster 2 from EPD genes

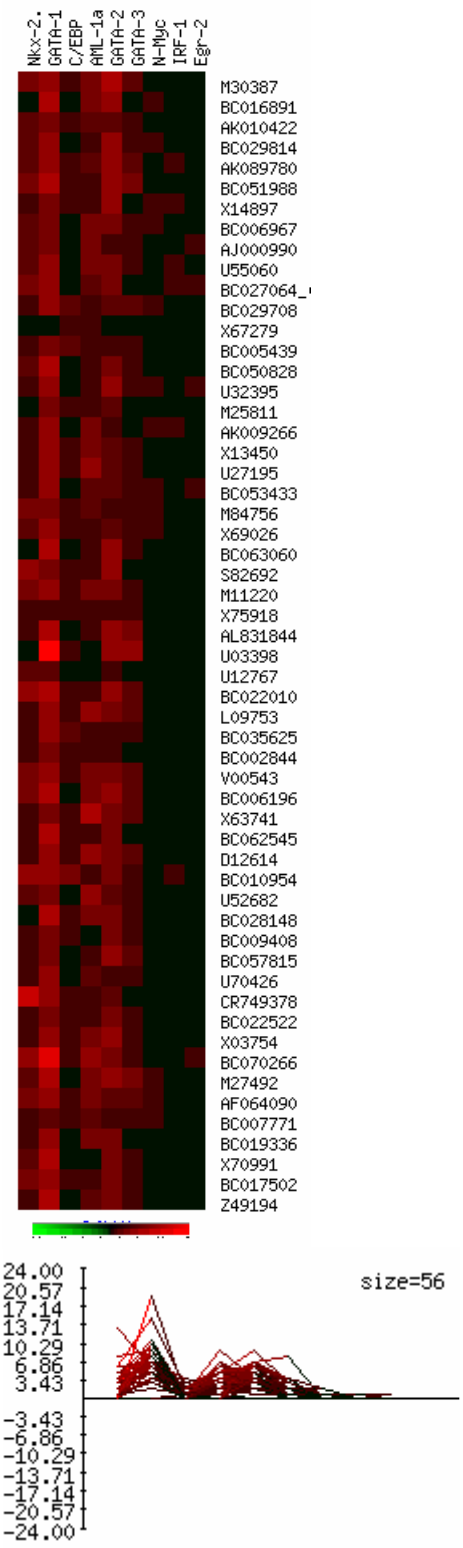


Figure 14. Cluster 1 from Microarray genes

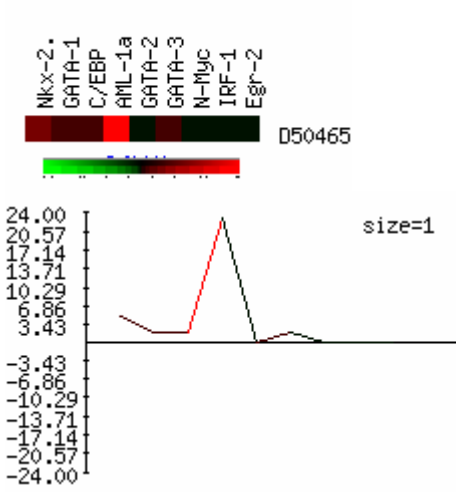


Figure 15. Cluster 2 from Microarray genes

5. DISCUSSIONS

5. 1. Overview of Significant Findings

This study was undertaken to computationally analyze transcriptional control elements and TFBS in lymphocyte development specifically in the expression of B cell and T cell genes.

There were 356 B cells specific Ig and 242 T cell specific upstream gene sequences TCR collected from RefSeq. Two predominant patterns were obtained from the Ig sequences and the TCR sequences using Gibbs Recursive Sampler (Table 1). These upstream sequences were isolated from regions enriched with a large number of germline variable (V) genes that are selectively recombined before expression. We are not sure that these are promoter sequences with in vivo transcriptional activities, but a biologically validated TF binding site Oct-1 was found within the motif identified from the Ig genes.

B cell specific and T cell specific TFBS were determined using the EZ-Retrieve software. We identified sets of non-overlapping B cell specific and T cell specific TFBS. We found 15 such B cell specific TFBS and nine T cell specific TFBS (Table 2). These results were arrived at by comparing EZ-Retrieve predicted TFBS from Ig and TCR genes with the EZ-Retrieve results of the Myers dataset. TFBS that have high probability of prediction are found in both TCR and Ig sequences while comparing with the Myers dataset. On the other hand, TFBS which have low probability are found in B cell genes but not in T cell genes and vice versa. The mutually exclusive TFBS in the B and T cell specific datasets were derived by EZ-Retrieve software as shown in Section 3 (Figure 6). As shown in Table 2, (A Union Im) is the same as A and (B Union Tm) is the same as B

which indicate that the B and T cell specific TFBS are unique sets of transcriptional control elements (TCEs) for these genes. These results imply that it is highly probable that B and T cells (at least at the level of Ig and TCR genes) employ these TFBS and the corresponding TFs for lineage specific developmental transcriptional programs.

It is not necessary that we obtain one hundred percentage reliable results using computational methods. In Table 2 (showing B and T cell specific TFBS) we have Egr-2 and Egr-3 both of which are not specific for B cell expression (51) and Oct-1 is a very common TFBS (50). p300 is not a DNA binding protein but still we predict the corresponding site to be a TFBS using EZ-Retrieve. NF-κB is found in both the B cell and T cell specific groups with different accession numbers.

Using two different algorithms, we found a set of genes with specific consensus motifs and tried to find any common genes that carried the same motifs in Gibbs Recursive Sampler and MEME predictions. We found that there were numerous common genes carrying various motifs predicted by the two algorithms (Table 3). A number of these genes have been implicated in lymphocyte development.

The two different programs are based on different algorithms. MEME is based on the expectation maximization algorithm and Gibbs Recursive Sampler is based on the Monte Carlo algorithm (explained in Methodology). Hence our predictions of common genes carrying the same motifs by the two methods indicate that these genes are likely to be involved in development of lymphocytes.

From the Microarray_T dataset, MEME and Gibbs Recursive Sampler identified one motif (MEME motif 1) where as in the Microarray_B dataset, both algorithms

identified one motif (MEME motif 2) that was in the reverse direction (one motif in '+' strand compared to the second in '–'strand).

Identification of co-regulated genes and their TFBS are key steps toward understanding transcriptional regulation. We performed EZ-Retrieve on all motifs obtained from MEME for the different datasets in order to obtain the TFBS and their locations and directions on motif, which could be biologically important.

EZ-Retrieve displays the TFBS predicted in the positive strands above the motif sequence while negative strands are displayed below. Some of these TFBS have been implicated as sites for binding immune specific TFs (*e.g.*., GATA, deltaE, E47, Oct-1) and maybe part of potential biologically meaningful transcriptional control elements (TCEs).

The alignment score between the Gibbs Recursive Sampler motifs and the different MEME motifs are higher compared to those obtained between MEME motifs aligned with each other. This indicates that MEME motifs are distinct patterns identified by the software. When the different MEME motifs were aligned with each other, we did not notice any significant sequence homology.

A statistical approach was implemented to identify a set of TFBS that could be employed to provide separation of genes specific for expression in B or T cells. For this analysis, the frequencies of TFBS were used as the distance measure to separate the two groups of genes. Then, we clustered these genes with the list of these informative TFBS as the metric for the clustering. We performed the statistical analysis and clustering on the EPD and microarray gene data sets. Clustering was performed with two sets of TFBS on each data set. The first TFBS set is above the cut-off value retrieved from the

BSS/WSS graph (statistical results) while the other was randomly selected from below the cut-off value in the BSS/WSS graph. It is expected that clustering which is performed using the TFBS above the cut-off will show better separation between B and T cell genes (in their respective clusters) as compared to randomly selected TFBS below the cut-off.

Separation takes place better with TFBS above the cut-off because of the high BSS/WSS value associated with them. This indicates that the TFBS above the cut-off have more distance between the sums of the squares of TFBS of the two groups of genes. This may cause better separation of B cell and T cell genes. The randomly selected TFBS below the cut-off have low values of BSS/WSS. This indicates that within the sum of the squares is greater thus the TFBS frequencies of the two groups of genes are closer and therefore the separation between B and T cell genes is not very good. We did not good separation of the B and T cell-specific genes in either cluster (in both data sets) with the “informative” TFBS (Figures 12 – 15). This is probably due to the very low BSS/WSS ratio thresholds (0.05 for EPD genes and 0.054 for microarray genes) that we had to use for separating the “informative” TFBS from the random ones.

We also performed MEME on randomly generated sequences of similar size to our test sequences but we could not get any motif because the E-values were high for such sequences. This negative result on random sequences confirms our motif prediction results from the B and T cell specific datasets.

5.2. Consideration of Findings in Context to Current Knowledge

It is difficult to discover information that is hidden in a long sequence of DNA. MEME and Gibbs Recursive Sampler software are good at retrieving short sequences called motifs. These motifs are very useful for studying transcriptional regulation.

If we have short sequence motifs, EZ-Retrieve can be utilized to find transcriptional control elements from these motifs. EZ-Retrieve is also useful at locating where different TFBS are present and their direction.

Statistical methods help to classify the different groups of genes with the help of K-means clustering.

Experimental methods for finding motifs or transcriptional factor binding sites are time consuming. However, bioinformatics methods are more useful at finding motifs and TFBS. These methods are frequently applied on different databases to find TFBS. In addition to effective laboratory assays, various computational approaches for detection of TFBS in promoter regions of co-expressed genes have been developed. Benefits of these methods are that it takes less time and resources over traditional laboratory methods.

6. CONCLUSIONS

Overview of Findings

We have used several computational methods in this project in order to analyze and identify biologically meaningful transcriptional control elements involved in lymphocyte development. While performing computational approaches like MEME, Gibbs recursive sampler, statistical analysis and k-means clustering on different DNA (promoter) sequences, we identify numerous biologically meaningful transcriptional control elements involved in lymphocyte development. For example, from B cell and T cell gene sequences, two predominant Gibbs pattern sequences were obtained. Different TFBS were obtained from EZ-Retrieve results for B and T cells and few of them are biologically important. In the case of B-cell, specifically important are Oct-1 and GATA-1. This informatics approach to detect transcriptional control elements may support the biologist studying transcriptional regulation that distinguishes B and T cell development.

Future Work

The future plan for this project is to perform statistical analyses on the significance of our predictions**.** This can be done by applying different statistical methods on the results and predicting the relevance of the result. Cross-validation is one of them, which can help to analyze the results of different software and predict better results. Another method is false positive prediction which can help to find the incorrect results which are predicted by the software. By applying other confidence measures we can modify our results such as controlling the P-value and E-value, which may help to find better motifs from the data sets.

In future analyses, it may be worthwhile to remove repeat sequences (using software such as RepeatMasker) before we run our upstream sequences through the transcriptional control element prediction algorithms. This may improve the validity of our prediction results even though some of these elements may actually reside and be transcriptionally active within repeat sequences.

7. REFERENCES

1. S. Qin, L.A. McCue, W. Thompson, L. Mayerhofer, C.E. Lawrence, and J.S. Liu. Identification of co-regulated genes through bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol* 2003 **21**, 435-9.

2. Y.J. Hu. Finding subtle motifs with variable gaps in unaligned dna sequences. *Comput Methods Programs Biomed* 2003 **70,** 11-20.

3.http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/P/Promoter.html
The_basal_promoter in transcription, Figure 1.

4. Gross P, Oelgeschlager T. Core promoter-selective RNA polymerase II transcription. *Biochem Soc Symp* 2006 **73,** 225-36.

5. Carmen Hernández-López, Alberto Varas, Rosa Sacedón, Eva Jiménez, Juan José Muñoz, Agustín G. Zapata, and Angeles Vicente. Stromal cell-derived factor 1/CXCR4 signaling is critical for early human T-cell development. *Blood(Immunobiology)* 2002 **99,** 546-554.

6. Eric Meffre, Rafael Casellas & Michel C. Nussenzweig. Antibody regulation of B cell development; Antibody regulation of B cell development. *Nature Immunology* 2000 **1,** 379 – 385.

7. Akashi et al. Lymphoid development from stem cells and the common lymphocyte progeniters. *Symposium 8.Immune Biology* 1999 **64,** 1-12.

8. Charles A. Janeway, Jr, Paul Travers.The development of B lymphocytes. *Immunobiology The Immune System in Health and Disease*, Current Biology/Garland Publishing Garland Publishing 717 fifth Avenue, New York, NY 10022 USA, 1994 5.1-5.32.

9. Haynes BF, Martin ME, Kay HH, Kurtzberg J.. Early events in human T cell ontogeny. Phenotypic characterization and immunohistologic localization of T cell precursors in early human fetal tissues. *J Exp Med* 1989 **169,** 1061-80.

10. Stephen Greenbaum and Yuan Zhuang. Regulation of early lymphocyte development by E2A family proteins. *Semin Immunol* 2002 **14,** 405-14.

11. Mary O'Riordan and Rudolf Grosschedl. Coordinate Regulation of B Cell Differentiation by the Transcription Factors EBF and E2A. *Immunity* 1999 **11,** 21-31.

12. Reimold AM, Ponath PD, Li YS, Hardy RR, David CS, Strominger JL, Glimcher LH. Transcription factor B cell lineage-specific activator protein regulates the gene for human X-box binding protein 1. *J Exp Med* 1996 **183,** 393-401.

13. Smith EM, Gisler R, Sigvardsson M. Cloning and characterization of a promoter flanking the early B cell factor (EBF) gene indicates roles for E-proteins and autoregulation in the control of EBF expression. *J Immunol* 2002 **169,** 261-70.

14. Nosaka T, Kawashima T, Misawa K, Ikuta K, Mui AL, Kitamura T.
.STAT5 as a molecular regulator of proliferation, differentiation and apoptosis in hematopoietic cells. *EMBO J* 1999 **18,** 4754-65.

15. Morgan B, Sun L, Avitahl N, Andrikopoulos K, Ikeda T, Gonzales E, Wu P, Neben S, Georgopoulos K. Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation. *EMBO J* 1997 **16,** 2004-13.

16. Yucel R, Kosan C, Heyd F, Moroy T. Gfi1:green fluorescent protein knock-in mutant reveals differential expression and autoregulation of the growth factor independence 1 (Gfi1) gene during lymphocyte development. *J Biol Chem* 2004 **279,** 40906-17.

17. Massa S, Junker S, Schubart K, Matthias G, Matthias P. The OBF-1 gene locus confers B cell-specific transcription by restricting the ubiquitous activity of its promoter. *Eur J Immunol* 2003 **33,** 2864-74.

18. Ellen V. Rothenberg and Tom Taghon. Molecular Genetics of T cell Development. *Annu. Rev. Immunol* 2005 **23,** 601-49.

19. Herblot S, Steff AM, Hugo P, Aplan PD, Hoang T. SCL and LMO1 alter thymocyte differentiation: inhibition of E2A-HEB function and pre-T alpha chain expression. *Nat Immunol* 2000 **1,** 138-44.

20. Guidos. Notch signaling in lymphocyte development. *Semin Immunol* 2002 **14,** 395-404.

21. Anderson MK, Weiss AH, Hernandez-Hoyos G, Dionne CJ, Rothenberg EV. Constitutive expression of PU.1 in fetal hematopoietic progenitors blocks T cell development at the pro-T cell stage. *Immunity* 2002 **16,** 285-96.

22. He X, He X, Dave VP, Zhang Y, Hua X, Nicolas E, Xu W, Roe BA, Kappes DJ.. The zinc finger transcription factor Th-POK regulates CD4 versus CD8 T-cell lineage commitment. *Nature* 2005 **433,** 826-33.

23. Jie Wen, Suming Huang, Svetlana D. Pack, Xiaobing Yu, Stephen J. Brandt, and Constance Tom Noguchi. Tal1/SCL Binding to Pericentromeric DNA Represses Transcription. *J. Biol. Chem* 2005 **280,** 12956-12966.

24. Patrick J. Daniels and Glen K. Andrews. Dynamics of the metal-dependent transcription factor complex in vivo at the mouse metallothionein-I promoter. *Nucleic Acids Res* 2003 **31,** 6710-6721.

25. Jen-Chywan Wang, Mika Kakefuda Derynck, Daisuke F. Nonaka, Daniel B. Khodabakhsh, Chris Haqq, and Keith R. Yamamoto. Chromatin immunoprecipitation (ChIP) scanning identifies primary glucocorticoid receptor target genes. *Biochemistry Proc Natl Acad Sci U S A* 2004 **101,** 15603-15608.

26. Silvia Smaldone, Friedrich Laub, Cindy Else, Cecilia Dragomir, and Francesco Ramirez. Identification of MoKA, a Novel F-Box Protein That Modulates Krüppel-Like Transcription Factor 7 Activity. *Mol Cell Biol* 2004 **24,** 1058-1069.

27. Stephen Greenbaum and Yuan Zhuang. Identification of E2A target genes in B lymphocyte development by using a gene tagging-based chromatin immunoprecipitation system. *Immunology* 2002 **99,** 15030-15035.

28. Lars Johan Gjestrum, Lars Petter Hansen and Petter Wilberg. C3a and LPS mediated gene regulation in human mast cells examined by microarray and PCR.Laboratory of Immunology, Institute of Oral Biology (IOB). Faculty of Dentistry, University of Oslo, 2003.

29.Li T, Chen YH, Liu TJ, Jia J, Hampson S, Shan YX, Kibler D, Wang PH. Using DNA microarray to identify Sp1 as a transcriptional regulatory element of insulin-like growth factor 1 in cardiac muscle cells. *Circ Res* 2003 **93,** 1202.

30. Brian C. Schutte, Joseph P. Mitros, Jennifer A. Bartlett, Jesse D. Walters, Hong Peng Jia, Michael J. Welsh, Thomas L. Casavant, and Paul B. McCray, Jr.. Discovery of five conserved β-defensin gene clusters using a computational search strategy. *Genetics* 2002 **99,** 2129–2133.

31. Voichita D. Marinescu, Isaac S. Kohane, and Alberto Riva. The MAPPER database: a multi-genome catalog of putative transcription factor binding sites. *Nucleic Acids Res* 2005 **33,** D91-D97.

32. Liu Y, Wei L, Batzoglou S, Brutlag DL, Liu JS, Liu XS. A suite of web-based programs to search for transcriptional regulatory motifs. *Nucleic Acids Res* 2004 **32,** W204-7.

33. K. T. Takusagawa and D. K. Gifford. Negative information for motif discovery. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge,MA 02139,USA.

34. Rongxiang Liu, Richard C. McEachin, and David J. States. Computationally identifying Novel NF-κB-Regulated Immune Genes in the Human Genome
Genome Res 2003 **13,** 654-661.

35.Shannan J. Ho Sui1, James R. Mortimer, David J. Arenillas, Jochen Brumm, Christopher J. Walsh, Brian P. Kennedy and Wyeth W. Wasserman. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res* 2005 **33,** 3154-3164.

36. Carl E Allen, Chi-ho Mak, and Lai-Chu Wu. The κB transcriptional enhancer motif and signal sequences of V (D) J recombination are targets for the zinc finger protein HIVEP3/KRC: a site selection amplification binding study. *BMC Immunol* 2002 **3,** 1-23.

37. Debraj GuhaThakurta, Lisanne Palomar, Gary D. Stormo, Pat Tedesco, Thomas E. Johnson, David W. Walker, Gordon Lithgow, Stuart Kim, and Christopher D. Link. Identification of a Novel cis-Regulatory Element Involved in the Heat Shock Response in Caenorhabditis elegans Using Microarray Gene Expression and Computational Methods 2002 **12,** 701-712.

38. Kim D. Pruitt and Donna R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 2001 **1,** 137-140.

39. Rouaïda Cavin Perier, Thomas Junier, Claude Bonnard and Philipp Bucher. The Eukaryotic Promoter Database (EPD): recent development. *Nucleic Acids Res* 2000 **28,** 307-309.

40. Robert Mansson, Panagiotis Tsapogas, Mikael Akerlund, Anna Lagergren, Ramiro Gisler, and Mikael Sigvardsson. Pearson Correlation Analysis of Microarray Data Allows for the Identification of Genetic Targets for Early B-cell Factor. *J. Biol. Chem* 2004 **279,** 17905-17913.

41. Arvind Raghavan, Rachel L. Ogilvie, Cavan Reilly, Michelle L. Abelson, Shalini Raghavan, Jayprakash Vasdewani, Mitchell Krathwohl and Paul R. Bohjanen. Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res* 2002 **30,** 5529-5538.

42. Shirin Khambata-Ford, Yueyi Liu, Christopher Gleason,, Mark Dickson, Russ B. Altman, Seraifim Batzoglou, and Richard M. Myers. Identification o Promoter Regions in the Human Genome by Using a Retroviral Plasmid Library-Based Functional Reporter Gene Assay. *Genome Research* 2003 **13,** 1765-1774.

43. Zhang, H., Ramanathan, Y., Soteropoulos, P., Recce, M., and Tolias, P.P. (2002). EZ-Retrieve: A web-server for batch retrieval of coordinate-specified human DNA sequences and underscoring putative transcription factor-binding sites. *Nucl. Acids. Res* 2002 **30,** e121.

44. TESS: Transcription Element Search System; http://www.cbil.upenn.edu/cgi-bin/tess/.

45. Trimothy L. Bailey and Michael Gribskov. Methods and statistics for combining motif match scores. *Journal of computational biology* 1998 **5,** 211- 221.

46. William Thompson1, Eric C. Rouchka2 and Charles E. Lawrence. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* 2003 **31,** 3580-3585.

47. ClustalW: http://www.ebi.ac.uk/clustalw/.

48. Dudoit S, Fridlyand J, Speed T . Comparison of discrimination methods for the classification tumors using gene expression data. *J Amer Stat Assoc* 2002 **97,** 77-87.

49. K-means clustering: http://ep.ebi.ac.uk/EP/EPCLUST/.

50. McCune RC, Syrbu SI, Vasef MA. Expression profiling of transcription factors Pax-5, Oct-1, Oct-2, BOB.1, and PU.1 in Hodgkin's and non-Hodgkin's lymphomas: a comparative study using high throughput tissue microarrays. *Mod Pathol* 2006 **10,** 1038.

51. Hui Shao, Dwight H. Kono, Ling-Yu Chen, Elyssa M. Rubin, and Jonathan Kaye. Induction of the Early Growth Response (Egr) Family of Transcription Factors during Thymic Selection. *J. Exp. Med* 1997 **185,** 731-744.

8. APPENDIX

Notes

Promoter: Nucleotide sequence of DNA where the sigma factor of RNA polymerase binds during transcription.

Promoter region: Specific initiation site of DNA where the RNA polymerase enzyme binds for transcription on the DNA.

RNA polymerase: An enzyme that catalyzes the formation of RNA macromolecules from DNA.

Transcription: The synthesis of mRNA, rRNA, and tRNA from a DNA template.

Transcriptional Factors (TFs): Eukaryotic proteins that bind to DNA and are responsible for binding the correct RNA polymerases to their correct promoters; proteins that bind DNA at a specific promoter site independently of RNA polymerases.

T cell Receptor (TCR): A receptor on the surface of a T cell that in association with either CD4 or CD8 is responsible for MHC-restricted antigen recognition; a heterodimer of two polypeptide chains that are anchored to the T cell membrane and contain immunoglobulin-like constant domains and amino-terminal variable domain.

T cell: T lymphocyte; lymphocyte cell that is differentiated in the thymus and is important in cell-mediated immunity, as well as in the modulation of antibody-mediated immunity.

T helper cell (TH): A class of T cells with CD4 markers that enhance the activities of B cells in antibody –mediated immunity; T lymphocytes that act as effector cells and interact with other T cells, B cells and macrophages to activate the immune response.

T Suppressor cell (TS): A class of T cells that produce cytokines that depress the activities of B cells in antibody-mediated immunity and other T cells and macrophages in cell mediated immunity.

Curriculum Vitae

Vandana Singh
6724 Caribou Court
Indianapolis, IN 46278

(317) 229-0212 (home)
(317) 331-5888 (mobile)
vandanaraisingh@yahoo.com

Education:

M.S.,Bioinformatics	Indiana University - Purdue University (IUPUI), Indianapolis	2006
M.S.,Chemistry	Banaras Hindu University – India	2000
B.S., Chemistry	Banaras Hindu University – India	1998

Work Experience:

Research Analyst, (July 2004 – present)
CAEC, LLC, Starkville, Mississippi

- Work part-time helping to analyze and process experimental data
- Used PowerPoint to create presentations

School Work Experience:

Laboratory Information Management System (Jan 2005 – May 2005)
Used Labware and LabVantage software to perform different analysis, tests on pain killers (Aspirin and Tylenol) to get the report

M.S. Thesis, (July 2004 – 2006)
Working on M.S. thesis titled, "**Computational Detection and Analysis of Transcriptional Control Elements in Lymphocyte Development**

Forestry Management System (Jan 2004 – May 2004)
Worked in a group to develop an application for collecting forestry data via the Web. This application used Oracle and ASP.NET.

Patient Information Tracking System (Sept 2003 – Dec 2003)
Lead a group of three students to design and implement a project for tracking patient data. This system was designed to allow doctors/nurses access to patient records via the Internet, schedule/cancel appointments and view medication history. The system used an Oracle database, XML and ASP.NET/C# technologies.

Face Recognition System (Jan 2002 – May 2002)
Developed application for face recognition using Artificial Neural Network (ANN) and MATLAB. The application was able to recognize 94% of faces feed into the system.

Publication/Honors:

Poster selected for presentation at the 13th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), June 2005, Detroit, Michigan.

Skills:

Programming Languages	C, XML and SQL
Databases	Oracle, MS Access
Bioinformatics Software	Microsoft Office 2000, MATLAB, MEME, BLAST, GIBBS Motifs Sampler, EPCLUST, EZ-Retrieve, Labware, Labvantage
Operating Systems	MS Windows (9x, 2000, XP), UNIX

References:

Available upon request